DOCTOR OF PHILOSOPHY

Tumour Localisation in Histopathology Images

Akbar, Shazia

Award date:
2015

UNIVERSITY OF DUNDEE

Tumour Localisation in Histopathology Images

by

Shazia Akbar

A thesis submitted in fulfillment for the
degree of Doctor of Philosophy

in the
School of Computing

July 2015

Contents

List of Figures

List of Tables

Declaration of Authorship

I hereby declare that I am the author of this thesis; that all references cited have been consulted by me; that the work of which this thesis is a record has been done by me, and that it has not been previously accepted for a higher degree.

Shazia Akbar

Declaration by Supervisor

I hereby declare that I am the supervisor of the candidate, and that the conditions of the relevant Ordinance and Regulations have been fulfilled.

Stephen J. McKenna

Acknowledgements

First and foremost, I would like to thank my supervisors, Prof. Stephen McKenna, Prof. Alastair Thompson and Dr. Lee Jordan for their ongoing support throughout my studies. This thesis is a combination of all of our efforts and I am deeply grateful for your guidance. Stephen, thank you for teaching me the qualities of a good researcher. I have learned a great deal about academic writing and critical thinking from you which will follow me through the remainder of my career. Alastair, thank you for your selfless dedication to both my personal and academic development. Lee, thank you for your invaluable expertise and wisdom. I could not have asked for better role models.

I would not have made it this far without the undying support of my family and friends, with particular thanks to my mother, Nusrat Akbar, who encouraged me to always do my best in every aspect of my life. I would not be the women I am today without her. Thank you to my siblings, Tahira, Sairah, Smerah, Saima, Fariad and Adil, for being there through the ups and downs of my PhD. I am also deeply grateful to my close friends, who gave me sound advice and were there when I needed them most. I truly am lucky to have such generous people in my life.

Thank you to the Computer Vision and Image Processing (CVIP) group at the University of Dundee for sharing your talent and wisdom with me throughout the years. I will cherish the fond memories. I would also like to thank our collaborators, Dr. Telmo Amaral (Culture Lab, Newcastle University), Dr. Colin Purdie (NHS Tayside, Ninewells Hospital) and Dr. Philip Quinlan (Advanced Data Analysis Centre, University of Nottingham), for their support and advice throughout this research.

I am grateful to the School of Computing staff who made it a joy to work at the University of Dundee throughout the years. It has been a pleasure to work with you.

The research presented in this thesis is funded by the Engineering and Physical Sciences Research Council (EPSRC) and the University of Dundee.

Abstract

Immunohistochemical (IHC) assessment in cancer research is important for understanding the distribution and localisation of biomarkers at the cellular level. However currently IHC analyses are predominantly performed manually, increasing workloads and introducing inter- and intra-observer variability. Automation shows great potential in clinical research to reduce pathologists' workloads and speed up cancer research in large clinical studies. Whilst recent advancements in digital pathology have enabled IHC measurements to be performed automatically, the acquisition of manual annotations of tumours in scanned digital slides is still a limiting factor. In this thesis, an automated solution to tumour localisation is explored with the aim of replacing manual annotations. As an exemplar, human breast tissue microarrays stained with estrogen receptor are considered.

Methods for automated tumour localisation are described with a focus on capturing structural information in tissue by adopting superpixel properties in a rotation invariant manner, suitable for histopathology images. To incorporate essential contextual information, methods which utilise posterior tumour probabilities in an iterative manner are proposed. Results showed pixel-level agreements between automated and manual tumour segmentation masks ($\kappa = 0.811$) approach inter-rater agreement between expert pathologists ($\kappa = 0.908$). A large proportion of disagreements between automated and manual segmentations were shown to correlate to minor discrepancies, inconsequential for IHC assessment. IHC scores extracted from automated and manual tumour segmentation masks showed strong agreements (Allred: $\hat{\kappa} = 0.911$; Quickscore: $\hat{\kappa} = 0.922$), demonstrating the potential of automation in clinical practice across large clinical trials.

Glossary

Adenocarcinoma - An invasive epithelial malignant neoplasm that has a glandular origin.

Allred - Scoring system which stratifies a breast cancer patient's estrogen receptor status into cancers that are likely to respond to hormone therapy [5].

Antibody - A large Y-shape protein produced by plasma cells that is used by the immune system to identify and neutralise pathogens such as bacteria and viruses.

Antigens - Any substance that causes the immune system to produce antibodies against it. An antigen may be a foreign substance from the environment, such as chemicals, bacteria, viruses, or pollen.

Benign - A condition, tumour, or growth that is not cancerous and therefore cannot spread to other body sites.

Biomarker - Measureable indicator of a biological state or condition.

Carcinoma - Cancer which originates in epithelial cell linings of organs like the breast.

Counterstain - A second stain added to a previously stained tissue sample to make cellular details more distinct.

Epithelial cells - Cells bound together in sheets of tissue called epithelia; epithelia line the cavities in the body.

Gleason - Grading system used to help evaluate the prognosis of men with prostate cancer.

Grading - A measure of cell appearance in benign or malignant tumours.

Histology - Brand of biology that deals with the microscopic examination of tissue.

Immunohistochemistry - The process of detecting antigens (i.e. proteins) in cells of a tissue section by exploiting the principle of antibodies binding specifically to antigens in biological tissue.

Immunopositive - A positive result observed on immunostaining for the target substance.

Immunonegative - A negative result (i.e. no staining) observed on immunostaining for the target substance.

In-situ - Cancerous cells which have not invaded through the basement membrane where the tumour is initially formed.

Intra-observer variability - Variation one observer experiences when observing the same material more than once.

Inter-observer variability - Variation between results obtained by two or more observers examining the same material.

Invasive - Cancerous behaviour in which malignant tumour has spread from its site of origin to other tissues.

Lymphatic vessels - Carry a clear fluid, lymph, away from the breast. Lymph contains tissue fluid and waste products, as well as immune system cells.

Lymphocyte - White blood cell that determine the specificity of the immune response to infectious microorganisms.

Malignant - Has the potential to grow and invade the surrounding tissue, causing harm to the patient.

Metastasise - To spread to another part of the body, e.g. via blood vessels, lymph channels etc.

Microtome - An instrument for cutting thin sections for microscopic study.

Neoplasia - The presence or formation of new, abnormal growth of tissue. e.g. fibroadenoma.

Organelles - Organised or specialised structures within a living cell.

Pathology - Study of disease, typically a branch of medicine that deals with the laboratory examination of samples from the body for diagnostic or forensic purposes.

Quickscore - Immunohistochemical scoring system (see Allred).

Sarcoma - Malignancy that starts in connective tissues such as muscle tissue, fat tissue, or blood vessels.

Stroma - The supportive tissue of an epithelial organ, tumour etc. consisting of connective tissues and blood vessels.

Tissue arrayer - Equipment for creation of tissue microarray blocks.

Tissue bank - Repository for human tissue intended for clinical or research purposes.

Tissue microarray - Consists of a paraffin block in which up to 1000 separate tissue cores are assembled in array fashion to allow multiplex histological analysis.

Tissue microarray spots - Circular cores of tissue, typically measuring 0.6mm, extracted and sliced from a tissue block. Represents a sample of the original tissue block.

Whole mount slide - A slice extracted from a tissue block undergone tissue preparation and placed on a glass slide for analysis. Represents a complete surface representation of the tissue within a tissue block.

List of Abbreviations

AUC	Area Under ROC Curve
BoS	Bag-of-Superpixels
CRISP	Contextual RISP
CRF	Conditional Random Field
DCIS	Ductal Carcinoma *In-Situ*
DNA	DeoxyriboNucleic Acid
ER	(O)Estrogen Receptor
ER+ve	Positive expression of estrogen receptor
ER-ve	Negative expression of estrogen receptor
HER2	Human Epidermal Growth Factor Receptor 2
HSV	Hue, Saturation, Value colour channel
IHC	ImmunoHistoChemistry
MLP	Multi-Layer Perceptron
PR	Progesterone Receptor
RISP	Rotation Invariant Superpixel Pyramid
RGB	Red, Green, Blue colour channel
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
RNA	RiboNucleic Acid
S-BoS	Spatial Bag-of-Superpixels
SLIC	Simple Linear Iterative Clustering
SVM	Support Vector Machine
TMA	Tissue Microarray

List of Symbols

The symbols denoted in this thesis are written as follows. Scalar variables: lower case (e.g. n); constant: upper case (e.g. N); vector: bold lower case (e.g. $\mathbf{x}$); matrix: bold upper case (e.g. $\mathbf{X}$).

For consistency, the following notation is used throughout this thesis.

N	Number of images in a dataset
$\mathbf{x}_n, n \in 1...N$	nth 2D image in a dataset.
$\mathbf{y}_n, n \in 1...N$	nth ground truth label in a dataset.
T	Number of context iterations
$\mathbf{p}_n^{(t)}, t \in 1...T$	Probability classification map for image $\mathbf{x}_n$ in iteration t.
κ	Kappa agreement.
$\hat{\kappa}$	Weighted Kappa-squared agreement.

Spin-context

M	Number of grid locations in a single image.
$\mathbf{x}_{mn}, m \in 1...M$	Image patch in image $\mathbf{x}_n$ positioned at location m.
$\mathbf{f}_{mn}$	Feature vector for location m in image $\mathbf{x}_n$.
g	Function which returns a spin-context context descriptor.

RISP, CRISP

Z	Number of superpixels in a given image.
$\mathbf{s} = \{s_1, ..., s_Z\}$	List of superpixels where s_z denotes a single superpixel.
$\mathbf{F} = \{\mathbf{f}_1, ..., \mathbf{f}_Z\}$	Set of superpixel features for superpixels defined in $\mathbf{s}$.
$H_z, z \in 1...Z$	Two-dimensional spatial bag-of-superpixels (S-BoS) histogram for superpixel s_z.
K	Length of codebook.
$\mathbf{C}$	Superpixel codebook containing K visual words.
v_z	Visual word for superpixel s_z
$c(\cdot)$	Function which returns the centre point of a superpixel
R	Radius of circular support window in BoS, S-BoS and (image-level) RISP, in pixels.
R_c	Radius of circular support window used to construct context-level RISP, in pixels.
L	Number of pyramid levels in RISP.
A	Exponential growth factor of number of annuli between RISP levels.
B	Number of bins used to model posterior probabilities in a context-level RISP.

Nested cross-validation

U	Number of folds in nested cross-validation.
V	Number of sub-folds in each fold of nested cross-validation.
S	A set of labelled data.
S_u	Validation set for fold u.
$\bar{S_u}$	Training set for fold u.
G_v	Validation set for sub-fold v.

List of Publications

- **Akbar S.**, Jordan L., Thompson A.M., McKenna S.J. Tumor Localization in Tissue Microarrays Using Rotational Invariant Superpixel Pyramids. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, New York, U.S. 2015.

- McKenna S.J., Amaral T., **Akbar S.**, Jordan L., Thompson A.M. Immunohistochemical Analysis of Breast Tissue Microarray Images using Contextual Classifiers. In *Journal of Pathology Informatics*. 2013; **4**:13.

- **Akbar S.**, McKenna S.J., Amaral T., Jordan L., Thompson A.M. Spin-context Segmentation of Breast Tissue Microarray Images. In *Annals of the BMVA*. 2013; (4), 1-11.

- **Akbar S.**, Amaral T., McKenna S.J., Thompson A.M., Jordan L. Tumour segmentation in breast tissue microarray images using spin-context. In *Medical Image Understanding and Analysis*, Swansea, U.K. 2012.

- McKenna S.J., Amaral T., **Akbar S.**, Thompson A.M., Jordan L. Immunohistochemical analysis of breast tissue microarray images using contextual classifiers. In *MICCAI Workshop on Histopathology Image Analysis*, Nice, France. 2012.

The work described in this thesis has also been presented at *2012 Google Anita Borg Scholar Retreat, Zürich, Switzerland*; *2012 British Computer Society (BCS) London Hopper Colloquium, London, UK*; and *2012 British Machine Vision Association (BMVA) Computer Vision Summer School, Manchester, UK*.

To my mother, Nusrat.

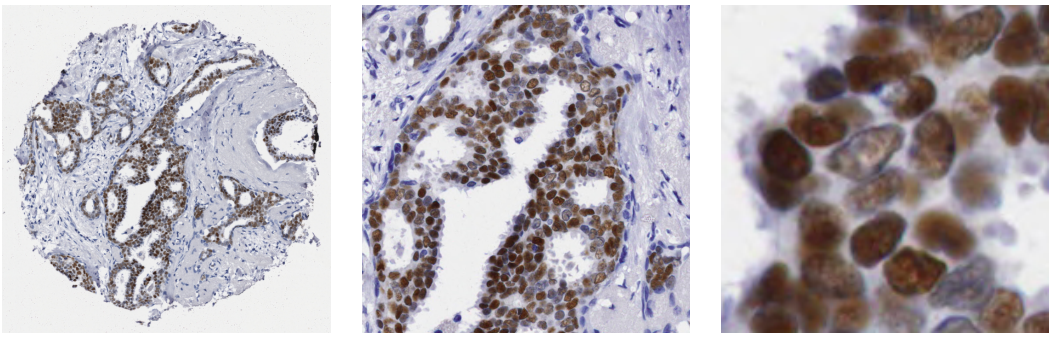
Chapter 1

Introduction

Cancer is a global health challenge. An estimated 14.1 million people were diagnosed with cancer in 2012, and this number is projected to rise to 26.4 million by 2030 [132]. To improve the treatment of, and survival from, cancer, a growing number of studies and clinical trials are enabling the collection of tissue samples [120]. In doing so, molecular analysis of DNA, RNA and protein expression can be performed.

Technological advances in digitisation of tissue slides and big data analysis have eased the burden of data handling, providing a means for collecting, sharing and storing vast amounts of data. However, the problem of analysing thousands of tissue samples for differences in protein expressions hence limiting understanding of cancer subtypes and individual treatments – persists. Currently, histopathological analysis of tissue must be performed manually by experts in pathology, thereby increasing pre-existing workloads. Manual analysis is also subject to inter- and intra-observer variability, resulting in non-standardised measures.

Digital slides, such as those shown in Figure 1.1, have introduced prospects of using image analysis techniques to improve manual analysis and increase clinical workflow. Advanced techniques in computer vision [58, 124] enable complex histological patterns to be modelled such that analysis of tissue sections can be performed automatically. In doing so, pathologists' workloads can be redirected to more difficult



(a) (b) (c)

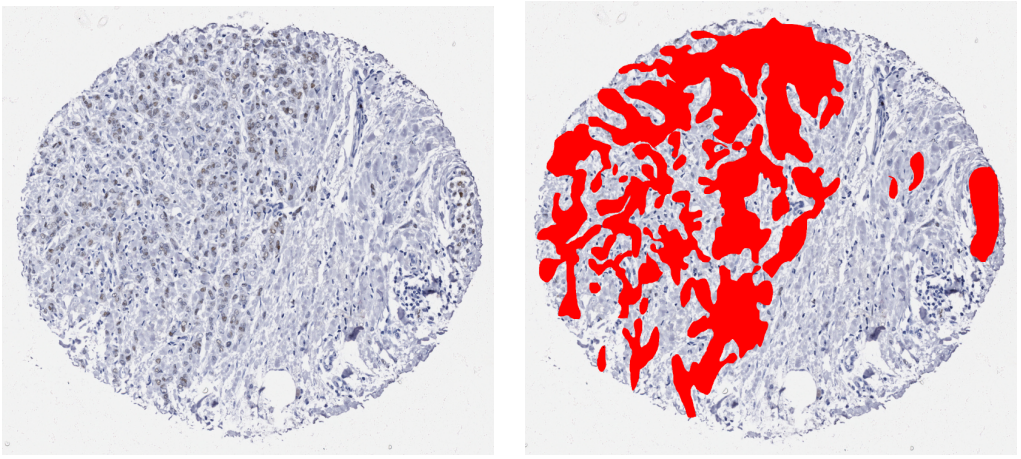
FIGURE 1.1: Estrogen receptor stained tissue microarray spot, scanned at x5 (a), x15 (b), and x40 (c) magnifications.

cases and analysis of large datasets in clinical trials can be increased, with potential for improving clinical workflow and throughput.

1.1 Immunohistochemistry

Immunohistochemistry (IHC), also sometimes referred to as protein expression profiling, is important for understanding the distribution and localisation of biomarkers. IHC enables observation of antigens in tissue sections by means of a specialised stain which binds to targeted antigens. In cancer research, IHC can provide prognostic data and predictive data informing treatment decision making.

Quantification of IHC biomarker presence is a challenging problem which currently requires expert knowledge to interpret biological material. Figure 1.1 demonstrates highly complex patterns and textures found in healthy and cancerous tissue at the microscopic level. Presence of estrogen receptor (ER) is shown in the form of a brown dye which is expressed in 80% of breast cancers [62]; the strength of the IHC stain can vary from cell to cell. To analyse multicellular structures e.g. glands, tissue is manually observed at various magnifications for essential contextual information. Any automated solution must also be adaptable to these changes.



(a) (b)

FIGURE 1.2: (a) Tissue microarray spot and (b) annotated tumour regions by expert pathologist.

1.2 Problem and motivation

In current practice, IHC measurements are acquired manually by visual inspection of tissue and localisation of tumours. Expert pathologists estimate the presence of cancer biomarkers in the form of IHC scores, computed from "ordinal scales" [5, 36]. However, this system lacks precision and introduces inter- and intra-observer variability [140]. An alternative approach is to utilise image analysis techniques which can potentially improve accuracy [92].

Currently the main bottleneck in histopathological image analysis, specifically IHC assessment, is the localisation of tumours. There are currently no guidelines in place for identification of tumours and it remains a challenging problem due to complex patterns and subjective analysis. Currently, in commercial software [11, 108], tumour boundaries are identified manually in the form of hand-drawn annotations (Figure 1.2). To move clinical research on a digital platform, annotations of this form must be acquired for each digital slide to be analysed. However gathering annotations for thousands of digital slides is time-consuming and not feasible given the practising pathologist's workload on a day-to-day basis. As such, an automated solution for localising tumour would be of great importance.

As an exemplar, breast tissue microarrays (TMAs) – circular paraffin-embedded sections of tissue – stained with ER (Figure 1.2(a)) are considered in this thesis. Whilst methods for automated analysis of TMAs are under development with the aim of speeding up pathology-based research of clinical materials, progress is slow due to the complex appearance of human tissue as well as artefacts (i.e. tissue folding, bubbles) introduced during TMA preparation. As the purpose of TMAs is to combine multiple tissue samples in one block, variability between samples is high. Tumour image analysis algorithms, when applied to TMAs, perform poorly due to lack of training examples.

In this thesis, techniques for modelling complex appearance of tumours are investigated. In the proposed methods, contextual information is captured in a rotation invariant manner, suitable for histopathological image analysis. To enable automated modelling of complex tissue structures, contextual information is explored in two forms:

1. To determine the extent to which surroundings contribute towards the classification of a single point in an image, context is captured from a circular support window. Properties within the window enable modelling of appearance, texture, geometry etc.

2. In the computer vision literature, contextual information has also been modelled in the form of posterior probabilities from learned classification maps [137]. In this thesis, alternative context descriptor representations are explored which capture the distribution of tumour posterior probabilities in a rotation invariant manner.

The automated methods described in this thesis are designed to identify cancerous structures without explicit labelling of tissue types (i.e. stroma, fat, epithelial cells). Furthermore, these methods can also be applied to nuclear, membrane and cytoplasmic stains, and therefore would be applicable to a range of biomarker research areas.

1.3 Contributions

In this thesis, techniques for automatic localisation of tumour in scanned digital slides are reported. Described methods were evaluated on a dataset of 32 ER-stained breast TMA spots and are summarised as follows.

Superpixels, compact groups of pixels, were used for the problem of tumour localisation in histopathology images. The intuition is that superpixels can indirectly model structure of tissue. For example, fibroblasts found in stromal regions result in more elongated superpixels than superpixels modelling epithelial cells. In this work, superpixel geometric, textural and appearance features were adopted for a novel representation called the Rotation Invariant Superpixel Pyramid (RISP).

In RISP, extracted superpixel features were quantised into superpixel visual words. To provide spatial information, annuli were positioned on a circular window, centred on the superpixel to be classified. Bag-of-superpixels (BoS) histograms were then computed per annulus resulting in a spatial bag-of-superpixels (S-BoS) histogram. It is shown spatial information in this form is essential for accurate tumour classification of superpixels. Furthermore to model visual words at multiple scales, the spatial pyramid [83] is extended to provide rotation invariance and incorporate superpixel properties. The RISP representation is a concatenation of S-BoS histograms in each level of the adapted pyramid structure. An experiment is reported which compares RISP with other rotation invariant superpixel representations, including superpixel autocorrelograms and a method proposed by Gorelick *et al.* [56] which captures context from *pixel-level* features in annuli. In all cases, RISP is shown to model complex patterns efficiently, resulting in superior performance.

Auto-context is a technique proposed by Tu and Bai [137] which models contextual information from learned classification maps in an iterative manner. In the past, auto-context has been successfully applied in the medical domain [68, 101]. Whilst auto-context has its merits, it is not rotation invariant and therefore not ideal for histopathology images. In this thesis an extension to auto-context, called spin-context, is described. In spin-context, auto-context is adapted to extract context locations from

points on equally-spaced circular rings thereby ensuring rotation invariance. To remove background interference, a further extension to spin-context is proposed which discards context locations outwith the region of a TMA spot. The removal of these context locations increased tumour classification accuracy in reported experiments.

To incorporate superpixels in the auto-context framework, a novel technique called Contextual RISP (CRISP) is proposed which models contextual information from superpixel posterior probabilities. In CRISP, *context-level* RISPs are proposed which model tumour distributions within annuli at multiple scales. An experiment was performed to compare CRISP with spin-context; CRISP was demonstrated to be superior with similar outcomes to RISP.

In addition to the proposal of image analysis techniques for tumour localisation, throughout this thesis automated and manual tumour segmentation masks are compared for clinical assessment. To measure inter-rater agreement between expert pathologists, a study was designed to compare hand-drawn tumour annotations. Annotations were gathered from two specialist pathologists and compared and assessed. To determine if automation can produce clinical measurements to the same standard as experts in pathology, IHC measurements were computed from automatically and manually-obtained segmentation masks, and an empirical evaluation was performed. It was found automation shows potential to replace manual input.

Pixel-level comparison of automated and manual segmentations is unsuitable for the current problem, as hand-drawn annotations are not accurate at the pixel-level. As such, an alternative evaluation technique was designed for categorising disagreements between binary segmentations. Disagreements are categorised into three types – Type 1, Type 2, Type 3 – with the focus being IHC assessment. Type 1 disagreements, minor discrepancies arising from lack of precision whilst drawing tumour boundaries, were shown to have little impact on extracted IHC measurements.

To summarise, the main contributions in this thesis are:

1. Extension of spin-context to reduce background interference.

2. Proposal of a Rotation Invariant Superpixel Pyramid (RISP) representation.

3. Proposal of context-level RISPs in a novel Contextual RISP (CRISP) framework.

4. Evaluation of automated and manual tumour localisation for IHC scoring, including a novel evaluation technique for categorising pixel-level disagreements.

1.4 Thesis outline

Chapter 1 Introduction: Description of problem and list of contributions.

Chapter 2 Background: Some background to clinical terms with reference to breast cancer statistics, histopathology, immunohistochemistry and tissue microarrays. The structure of breast cancer in histopathology images is discussed.

Chapter 3 Tumour Image Analysis in Digital Histopathology: A review of previous work in tumour segmentation and classification. Related work in cell classification, automated grading and, lobule and gland segmentation is explored.

Chapter 4 Manual Tumour Localisation: A study is reported, comparing annotations retrieved from two trained pathologists. Inter-rater agreement is reported and pixel-level disagreements are evaluated in more detail.

Chapter 5 An Extension and Evaluation of Spin-Context: A description of auto-context and spin-context with a review of related work in the medical domain. An extension of spin-context is described which eliminates background interference. Results are reported for multiple iterations of spin-context, and a comparison is made between spin-context and auto-context.

Chapter 6 RISP: Rotation Invariant Superpixel Pyramid: An introduction to superpixels and their applications, including a review of related work. The proposed RISP representation is described and results are reported comparing RISP with other superpixel representations. The CRISP framework is described and compared to RISP.

Chapter 7 Automated Tumour Localisation: Clinical Impact: A study comparing automated segmentation masks produced by RISP and manually hand-drawn segmentations. Agreements are reported for pixel-level evaluation, IHC scoring and ER treatment decision-making.

Chapter 8 Discussion and Conclusions: Summary of the main findings in previous chapters and comparison of the proposed approaches with inter-rater agreements. Limitations and potential of proposed methods are discussed.

Chapter 9 Recommendations: Outline of future directions for this research.

Chapter 2

Background

2.1 Breast cancer

Breast cancer is the most common cancer in women in the UK, with 1 in 8 women at risk of a diagnosis in their lifetime [24]. With medical and scientific advancements in cancer research, women diagnosed with breast cancer are now twice as likely to survive the disease for at least ten years when compared with patients diagnosed forty years ago. Despite this, nearly $12,000$ women die every year as a result of breast cancer in the UK. Latest statistics show over 90% of women diagnosed with breast cancer at the earliest stage survive their disease for at least five years. Relative survival in women ranges from 99% (stage 1) to 15% (stage 4) for patients diagnosed between 2002 and 2006 [24].

Various studies have been performed in the last decade to identify breast cancer risks and prevent development of future cancers. To date, several guidelines have been developed concurrent with this goal. Women over the age of 50 are more likely to develop breast cancer than men or young women. Family history and genetics also influence breast cancer risks. In addition, cancer risks vary considerably between countries and racial groups [81]. Lifestyle factors have also been shown to influence the risk of developing breast cancer, particularly intake of alcohol. Women who consume two to five drinks daily have about two times the risk of developing breast cancer

when compared with women who don't drink alcohol [9]. Obesity increases risk of postmenopausal breast cancer by up to 30% [24]. Other lifestyle factors including physical activity and eating habits further reduce breast cancer risks.

In the early stages of breast cancer, the most common treatment is surgical intervention however other treatments such as radiotherapy, chemotherapy and hormone therapy are also used to reduce the chance of recurrence. Each patient diagnosed with breast cancer undergoes a series of assessments performed by specialists including pathologists, oncologists, radiographers, surgeons and nurses. The clinico-radiological and pathological parameters combined with the general health of the individual patient determines the treatment recommendations.

After diagnosis, 1 in 5 women in the UK have a recurrence of their breast cancer within 10 years [95]. Given these recurrence rates and the growing number of women diagnosed with breast cancer, large breast cancer trials are vital to improve future treatment and often include the need to build tissue banks which facilitate breast cancer research worldwide. As an example, the Tayside Tissue Bank operating in Ninewells Hospital, U.K., contains in the region of $100,000$ freshly frozen tissue samples from around $10,000$ individual patients.

2.2 Cancer development in the breast

The female breast is mainly composed of lobules, ducts and stroma [109], as shown in Figure 2.1. Lobules are milk-producing glands which extend from ducts that carry the milk from the lobules to the nipple-areolar complex. Lobules and ducts are lined with myoepithelial and epithelial cells as shown in Figure 2.2, which are encased within stroma, consisting of fatty and connective tissue.

Cancer cells develop as a result of ungoverned growth of cells, and in the breast these typically originate at epithelial cells in ducts and lobules. When tumour develops uncontrollably, healthy breast structures are destroyed. If left untreated, cancer cells can metastasise and spread through lympho-vascular spaces to other parts of the body.

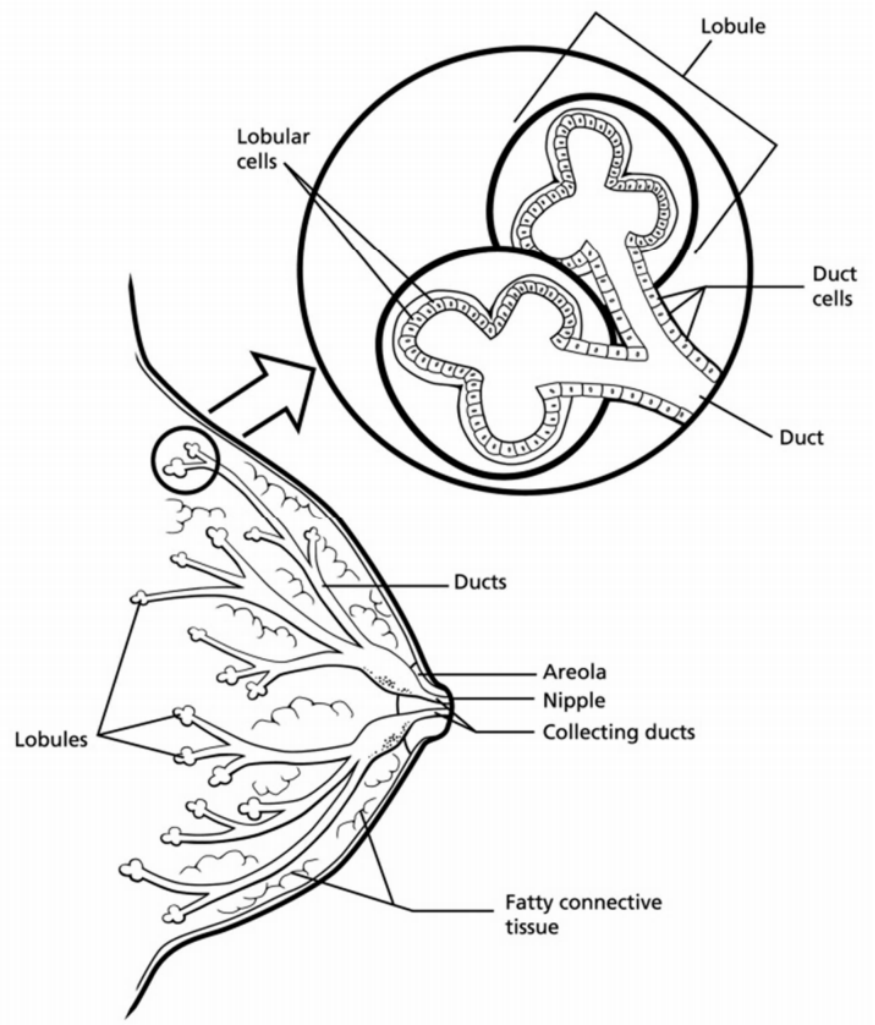


FIGURE 2.1: Overview of breast structure. Used with permission of American Cancer Society.

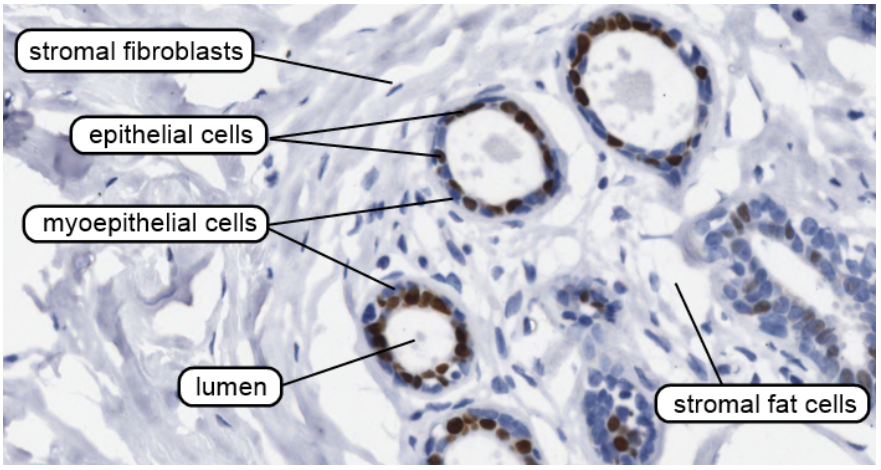


FIGURE 2.2: Histopathological structure of glands and lobules.

Types of breast cancers are categorised into two broad subgroups: *in-situ* and invasive. *In-situ* breast neoplasia is evident as cancer cells contained within the basement membrane of the breast ducts or lobules. *In-situ* types are unable to spread to other body sites and instead expand the ductal or lobular unit from which they originate. The most common *in-situ* breast cancer is ductal carcinoma *in-situ* (DCIS). Invasive breast cancer is cancer that spreads outside the basement membrane of the lobule or duct into the breast tissue. Invasive cancer cells infiltrate the connective tissue in the breast and can therefore spread via the lymphovascular channels to various structures including lymph nodes, and beyond. Once cancer spreads beyond the breast to develop metastasis, it is essentially incurable (though it remains treatable to prolong life).

2.3 Histopathology

In clinical medicine, histopathology refers to the microscopic examination of stained tissue biopsies in order to study disease. The process of creating stained slides is described in Section 2.4. Histopathology is the gold standard for tissue diagnosis and thereby confirms clinical suspicions enabling treatment to proceed and also provides data to inform treatment of neoplasic diseases and the patient's prognosis. Neoplasia is defined is an "abnormal mass of tissue, the growth of which exceeds and is uncoordinated with that of normal tissues, and persists in the same excessive manner after cessation of the stimuli which evoked the change" (R.A. Willis, British Oncologist). This type of behavioural growth is termed *malignant*. In contrast, sometimes tumours may grow locally and not spread to other areas of the body and are termed *benign* [109].

The textural appearance of cancer in stained slides differs depending on the body site from which it originates and the rate of cancer development. Figure 2.3 shows image patches extracted from estrogen receptor (ER) and haematoxylin (H) stained breast tissue slides. Images are shown for stromal, fat, lobular and glandular tissues found in the breast. Whilst some structures are easily distinguishable such as fat, many of

the other structures share similar textural and architectural appearances. For example, lobules have a similar structural appearance to ducts, particularly along the wall lining.

Cancer development within the breast introduces further complexities. In the early stages of breast cancer, cancer cells appear similar to epithelial cells and become abnormal with further development. In low grade DCIS (Figure 2.3(e)), nuclei are uniform in size and shape and are considerably smaller than in high grade DCIS (Figure 2.3(f)). However in both cases, cancer cells remain within the walls of the duct basement membrane and are therefore *in-situ*. Figures 2.3(g) and 2.3(h) show cases of invasive breast cancer. Notice that cancer cells have spread to the surrounding tissue. Walls of ducts and lobules are no longer visible, and the appearance and positioning of cancer cells are more irregular in size and shape than in DCIS.

2.4 Tissue preparation

In order to produce a translucent slice of tissue which can be observed under a light microscope, a series of preparation phases (fixation, tissue processing, embedding, sectioning) must be performed [109]. Prior to tissue processing, a biopsy is obtained from the patient, which is transported to a pathology laboratory in sealed containers to preserve the tissue. The biopsy then undergoes the following procedures sequentially.

Fixation: In the initial stage, the tissue is preserved in a steady state to undergo further preparative procedures. Fixation arrests autolysis (cell death) and bacterial decomposition, and stabilises the cellular and tissue structures. The use of chemical reagents such as formalin is considered the primary method of fixation.

Tissue processing: To remove water from the tissue, a dehydrating agent such as methylated spirit is applied. A clearing agent such as xylene acts as a link between the dehydrating agent and wax. It is common practice to start with a dilute clearing or dehydrating agent (70%) and gradually increase the concentration, as the use of absolute alcohol alone affects tissue adversely and results in increased shrinkage of the tissue.

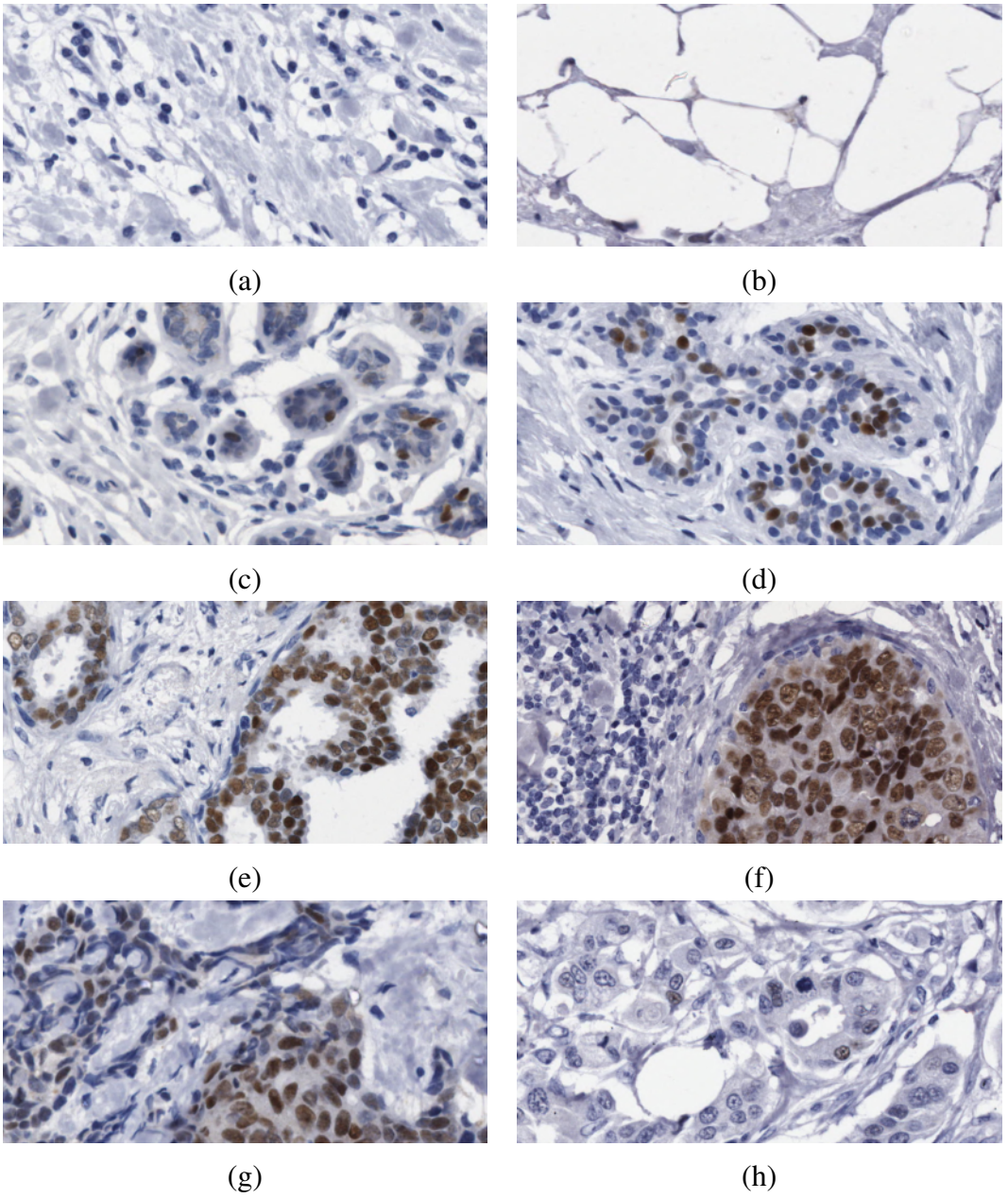


FIGURE 2.3: Image patches of estrogen receptor (ER) and haematoxylin (H) stained histology slides. Images are shown for stroma (a), fat (b) and healthy epithelial cells (c and d). Cancerous breast tissue is also shown for low (e) and high (f) grade ductal carcinoma *in-situ* and invasive breast carcinoma (ER rich (g) and ER poor(h)).

Embedding: During embedding, the tissue is positioned in a manner such that the maximum amount of diagnostic information can be obtained from the section. The most commonly used embedding medium is paraffin wax which has a consistency similar to that of tissue. To cool the wax, an electrically-controlled chilled area allows the pathology staff to orientate the tissue manually prior to complete solidification of the wax on a large refrigerated plate.

Sectioning: The final stage of tissue construction is sectioning, in which the tissue is cut into thin slices and mounted onto a piece of glass. With breast tissue, staff will typically slice the tissue to a width of 4 μm using a microtome. The resulting sections are sufficiently translucent to allow light to easily pass through the tissue. In order to place a tissue section onto microscope slide, it is then placed into a waterbath at a temperature just below the melting point of the wax. This allows the section to be "relaxed" out of the tissue before it is placed onto a microscope slide.

The result is a slice of tissue on a piece of glass which can be preserved for a prolonged period of time. The latest advancement of technology also allows pathologists to scan prepared slides at high resolutions using specialised scanners, thus allowing slices to be stored in digital form. In addition to ensuring prolonged storage of tissue samples, digital slides also allow pathologists to collaborate on a national and international level by sharing content online. The digitisation of slides has also introduced prospects of automated analysis of tissue to reduce pathologists' workloads, and potentially improve accuracy and reproducibility of pathologists' interpretations [124].

It is important to note that the method of preservation described in this section can lead to artefacts in the tissue introduced at various stages of tissue processing. For example, bubbles may appear on the slide if any pockets of air are trapped during sectioning. As such, both digital and physical slides may contain artefacts which can obscure or alter the true structure of the tissue.

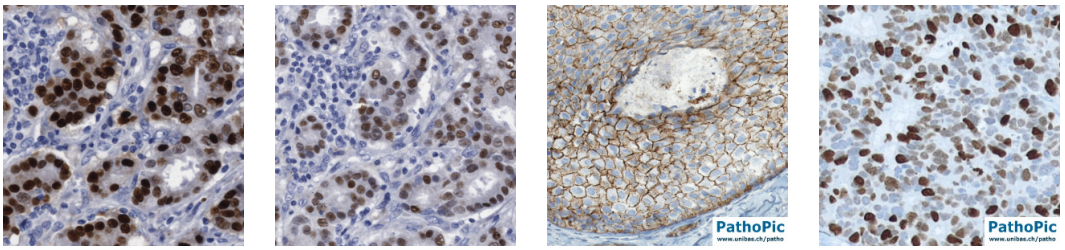
2.4.1 Stains

After preparation, the tissue slice is almost transparent and therefore requires a dye in order to be able to observe the structures of the tissue. The most common stain used in pathology is Haematoxylin and Eosin (H&E) which stains basic structures including cytoplasm, nucleus, organelles and extra-cellular components. The nuclei of cells are stained blue (H), and other components in pink/red (E). H&E generally gives an overview of the tissue. Other specialised stains are used if additional information is needed to provide a more detailed picture, for example to differentiate between two morphologically similar cancer types. Some examples of specialised stains for immunohistochemical tests for breast cancer are described in the following section.

2.5 Immunohistochemistry

Immunohistochemistry (IHC) was first reported in 1941 for detecting antigens in tissue sections by means of specific antibodies [34]. Since then the number of tests have grown considerably; IHC now has applications in diagnosis, prognosis, therapeutic decision making and studies of pathogenesis [122]. Within a tissue section subjected to IHC, cells that express the antigen-antibody reaction are termed immunopositive and appear stained. Immunonegative cells are not stained as a result of this reaction and are only visible as a result of the counter-stain e.g. H. The most common IHC stains used in breast cancer cases are estrogen receptor (ER), progesterone receptor (PR), HER2 and Ki-67 (Figure 2.4).

The presence of estrogen and progesterone hormone receptors has been shown to be an important prognostic and predictive biomarker in breast cancer [62]. ER and PR are nuclear stains and show their presence in the form of a positive signal, typically a brown stain (Figure 2.4). 80% of breast cancer cases express ER and 67% PR, however both receptors may also occasionally be found in healthy epithelial cells.



ER PR HER2 Ki-67

FIGURE 2.4: ER, PR, HER2 and Ki-67 immunohistochemical stained breast tissue.

HER2 is a cell membrane stain and is typically used to assess prognosis and to determine suitability for trastazumab therapy. An over-expression of HER2 protein in the breast occurs in 13.3% of invasive breast cancers [120].

Ki-67 is a nuclear protein which is present in all stages of the cell cycle but not in resting G_0 cells [53]. It provides the means to determine the growth fraction of a given cell population. Studies have shown Ki-67 is present in a wide range (2-80%) of breast cancer cases [55].

2.5.1 Immunohistochemical scoring

The IHC scoring of a tissue section refers to the process of quantifying cells which exhibit positive staining for a specific antibody. IHC scoring plays a key role in oncology to help characterise tumours and provide prognostic and predictive data [31]. The usage of IHC scoring for determining treatment is discussed later in this section.

Prior to IHC scoring, tumour regions are identified within the tissue slice so that healthy regions are discounted from resulting scores. This ensures scores reflect only abnormal tissue. Various scoring systems are available to pathologists; the type of scoring system adopted depends on personal preference and laboratory guidelines. In this thesis, Quickscore [36] and Allred [5] scoring systems are considered. Quickscore is used in clinical research at Ninewells Hospital, U.K., whereas Allred is a more widely used scoring system.

The Quickscore scoring system proposed by Detre *et al.* [36] was designed to be used to study the biology of breast cancers in a large number of sections and to determine the clinical implications of hormone treatment. The additive Quickscore is calculated by summing the stain intensity score ranging from 0 to 3 (0: negative, 1: weak, 2: moderate, 3: strong), with the percentage of positively IHC stained cells within tumour ranging from 1 to 6 (1: 0-4%, 2: 5-19%, 3: 20-39%, 4: 40-59%, 5: 60-79%, 6: 80-100%). When intensity and proportion scores are summed, this results in a Quickscore between 1 to 9. When there is negative IHC staining, the Quickscore defaults to 0. The Allred scoring system [5] adopts a different proportion scale ranging from 1 to 5 (1: 0-1%, 2: 1-10%, 3: 10-33%, 4: 33-66%, 5: 66-100%). The final Allred score is computed as in Quickscore, resulting in scores ranging from 0 to 8.

The IHC score for a specific biomarker is commonly used to determine whether or not a patient would benefit from treatment. For example if visual inspection of tissue shows strong presence of ER, then hormone therapy may benefit the patient by lowering estrogen in the body. The question arises as to what is the "correct" cut-off mark to determine immunopositivity (+ve) or immunonegativity (-ve). For ER, an Allred score > 2 (equivalent to USCAP [67] 1% cut-off) is termed ER positive (ER+ve); all scores ≤ 2 are ER negative (ER-ve) [5]. In Quickscore, a cut-off mark > 3 is commonly utilised, where ≥ 3 is ER+ve [36]. It is important to note that the cut-off marks defined here are not standardised. In the pathology literature, cut-off marks differ between biomarkers and scoring systems.

A key problem in manual IHC assessment such as scoring is inter- and intra-observer variability which occurs due to human error and differences in opinion when interpreting biological materials. Intra-observer variability refers to the variability of a single observer's score for a particular patient, whereas inter-observer variability is the variability amongst multiple observers. *Intra-observer variability* refers to a pathologist's disagreement with his/her own observation (i.e. made on multiple independent occasions). *Inter-observer variability* refers to disagreements between two or more pathologists. These types of differences effect the reliability of resulting measures in IHC tests. Typically, in cases whereby scores differ, pathologists either come to an

FIGURE 2.5: A tissue microarray of estrogen receptor stained breast tissue.

agreement through discussions, or multiple scores are retrieved from pathologists and an average is calculated.

2.6 Tissue microarrays

Tissue microarrays (TMAs) are constructed by extracting core samples from paraffin-embedded tissue specimens of multiple patients and transferring them to a single multicore paraffin block. An example of a TMA is shown in Figure 2.5. A TMA "spot" refers to a single section extracted from a single core sample; a slide from a TMA will contain multiple "spots". TMA spots are small circular sections of tissue, typically measuring 0.6mm in diameter. The preparation of TMAs is identical to the process described in Section 2.4 and is therefore subject to the same artefacts. Primarily, TMAs are used in clinical research for high-throughput molecular analysis to analyse various types of cancers. They enable preservation of tissue in the original block and maximise research-use of tissue whilst not compromising diagnostic uses in future. However, there is a strong potential for TMAs to also be used in clinical practice [134].

The technique for constructing TMAs was first reported in 1986 by Battifora [17] who described a "sausage-block" method in which he wrapped different tissue around a

small intestine which was embedded in a paraffin block. Wan *et al.* described the first array format for organising TMAs on a block effectively, in 1987 [146]. In order to separate each tissue slice in the block, Konenen *et al.* proposed a further method for rapid and accurate construction of tissue microarrays [78]. Today, hundreds of TMA spots can be organised on an indexed array block which can be easily referred to at a later date.

For the extraction of TMAs, a tissue microarrayer is required which has two hollow needles and a block holder that operates on a manual basis. The tissue microarrayer cuts TMA cores at designated locations identified by pathologists and places them into an empty paraffin block called the recipient block. Cores are arranged in a grid-like pattern in the recipient block. Sections from the recipient block are cut using a microtome, and are then mounted onto a glass slide for pathological analysis. The maximum number of 0.6mm spots is about 600 for a standard glass microscope slide.

In breast cancer research, TMAs are primarily used to evaluate prognostic and predictive biomarkers [14], which has become a mandatory step of the research pathology workflow [125]. TMAs are ideal for these conditions as they guarantee identical experimental conditions [112] and permit rapid assessment of individual molecular markers on patient cohorts [145]. TMAs also allow the preservation of patients' archival tissue for verification of clinical diagnosis and future diagnostic evaluation [97], essential for clinical research.

One of the main concerns about TMA analysis is that a small tissue core may not represent an entire tumour region in the donor block. Nevertheless, many studies have shown the use of TMAs is an economical replacement for whole section analysis of breast biomarkers [134]. A study validating TMA technology for immunohistochemical assays showed that in 95% of cases, two core sections were comparable to the results achieved from a whole tissue section [23]. Parker *et al* [111] similarly showed 96% agreement between whole sections and TMAs for estrogen receptor.

In this thesis, ER-stained breast TMAs are utilised as a clinically relevant exemplar

for the task of tumour localisation. Methods reported and conclusions drawn in subsequent chapters are also applicable to whole mount slides and are designed to scale with ease.

Chapter 3

Tumour Image Analysis in Digital Histopathology

3.1 Introduction

With the recent advancement and cost-effectiveness of digital scanners, tissue histopathology slides can now be scanned and stored in digital form. A digital slide can be viewed at 40x magnification enabling detailed observations to be performed at the cellular level. Furthermore, digital slides are not subject to tissue degradation and can therefore be archived and retrieved easily. Recent sophisticated imaging and analysis techniques have introduced the prospect of histopathological image analysis to ease or aid manual analysis of digital slides. By cutting time that is spent diagnosing normal tissue, pathologists can adjust their workload to more difficult cases [124] and analysis in large clinical trials can be performed with ease. By ensuring repeatability, automation diminishes the problem of inter- and intra-observer variability [142].

In histopathology, image analysis algorithms have been applied to the problem of disease prognostics, disease grading, protein and gene expression, and much more. In

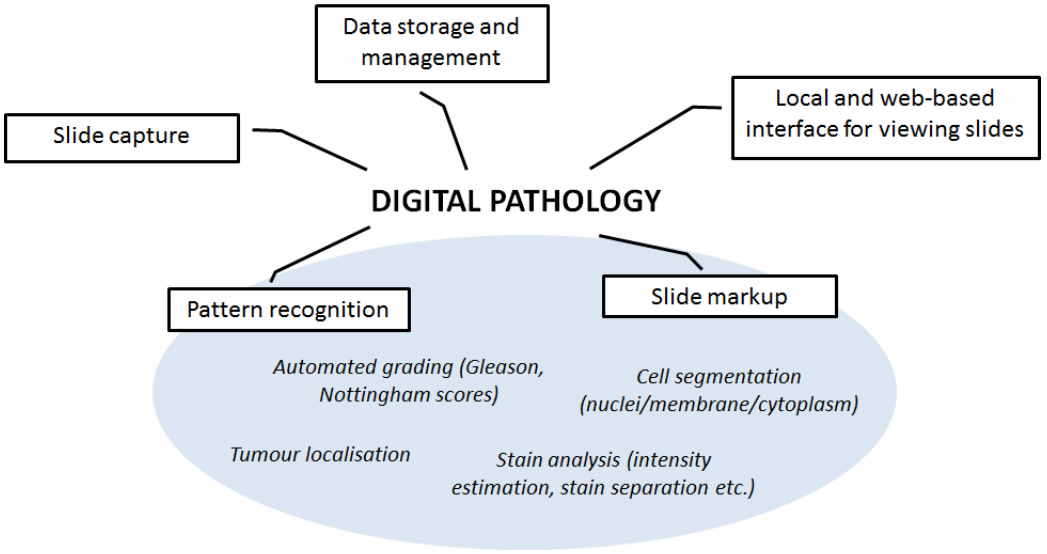


FIGURE 3.1: Overview of features available in digital pathology applications.

general, automation can be helpful for certain tasks, such as cell annotation, as a human expert cannot be accurate to the pixel level and can easily overlook cells. However, in order to solve specific tasks in histopathology, algorithms must be tailored towards their intended purposes. "If the ultimate objective of the Computer Aided Diagnosis algorithm is, for instance, cancer grading, perfect segmentation of histological structure may not guarantee perfect grade-based classification." [58]

Currently, algorithms developed for histopathological image analysis indicate positive outcomes for future uses of such applications in clinical practice. However, they have yet to reach the accuracy of a human expert. Therefore replacement of current human expertise in this field is unreasonable in the short- or medium-term [92]. Instead, image analysis methods complement the role of the pathologist and learn from human expertise. For example, Mercan *et al.* [96] integrated the pathologist into their system by learning from both expert navigation behaviour and image features to automatically identify regions of interest in whole mount slides. In terms of future prognosis, imaging techniques are providing ways to assist the pathologist in making accurate diagnosis and identifying morphological features related to prognosis.

Digital pathology in the commercial industry has developed considerably in the last decade with a range of applications from data storage to pattern recognition software.

Figure 3.1 shows an overview of some of the available features in current commercial software [10, 35, 84, 85, 108]; however they are by no means limited to these applications. In terms of TMA-specific operations, many applications offer TMA dearraying software [10, 66] to automatically extract TMA spot images from scans of TMA array blocks. In Figure 3.1 image analysis features which have grown and developed over the last decade are highlighted in blue. These typically stem from slide markups (i.e. annotations) and trained models from pattern recognition algorithms.

In this chapter, image analysis algorithms in digital histopathology are explored for the purpose of tumour analysis. This work ranges from cell-level analysis (Section 3.2) to models of intermediate structures incorporating broader context (Section 3.3 - 3.4). Techniques adopted to model tumour are described with reference to work in the academic and commercial sectors. In particular, IHC assessment using these methods is discussed.

3.2 Cell segmentation

The detection and segmentation of cells has been widely researched in several fields including histopathology, cytology and microscopy for many years. By identifying individual cell nuclei, membrane or cytoplasms, cells can be classified on the basis of antigens which mark specific cellular features.

Level sets is a common technique used for cell segmentation [63, 127, 135, 152]. The property of closed contours is appropriate for modelling individual cells. Level set methods are also efficient for modelling topological changes such as merging and splitting cells. In recent work, Nath *et al.* [103] studied three level set techniques and concluded that an optimised version of the N-level set formulation [156] was ideal for cell segmentation. Qi *et al.* [121] modified the level set energy function to compare neighbouring contours, thereby separating overlapping cells. Yu *et al.* [155] described a level set technique which propagated faster in regions of brighter image intensities in flourescence imaging.

Other methods use prior knowledge about the elliptic shape of cells to improve localisation. Ni *et al.* [107] developed a voting algorithm in the shape of an ellipse to detect symmetrical visual patterns in histopathology images. Chomphuwiset *et al.* [30] used an ellipse fitting model to detect cell nuclei which were then classified based on mean colour priors.

The above techniques assume prior knowledge about the shape of cells to be segmented. An alternative approach is to learn cell shape for more accurate segmentation. In early work, Thiran *et al.* [133] performed a series of morphological operations to estimate the shape, size and texture of cells during training. More recently, Arif and Rajpoot [12] used boundary points from k-means segmented objects in the manifold learning framework to learn the shape of cells during classification.

In commercial software, several tools are available for quantification of cells which exhibit positive IHC expression. IHC-MARK [108] by Oncomark automatically segments IHC positive cells via watershed segmentation [52]. Similarly, Definiens Tissue Studio provides a built-in tool for IHC cell analysis, applicable to TMAs [21]. Other software packages including Aperio ImageScope [11] and Indica Halo [65] offer similar services for IHC assessment.

3.2.1 Lymphocyte cell segmentation

In addition to cell analysis, lymphocyte infiltration has also been shown to be a strong indicator of cancer development [3]. In a pattern recognition contest for lymphocyte counting in histopathological images [58], Kuse *et al.* [80] produced promising results by distinguishing cells via a mean shift based clustering and HSV thresholding technique. Lymphocytes were classified based on textural features extracted from identified cellular contours.

Chomphuwiset *et al.* [30] argue that it is difficult to filter lymphocytes from histopathological images as they share similar properties to epithelial cells. To solve this problem, Panagiotakis *et al.* [110] proposed a model for segmenting three classes: stroma,

cell nuclei and lymphocyte nuclei. The proposed method adopts the maximum likelihood principle to classify image sites and an Expectation-Maximisation (EM) algorithm to estimate parameters. Fatakdawala *et al.* [47] also proposed a model which uses EM to automatically initialise active contours to segment lymphocytes in HER2+ breast tissue.

Whilst existing models show cell segmentation can be performed accurately with appropriate elliptical or shape models, the main difficulty arises in the classification of cells. At the cellular level, some properties e.g. cell shape and size, are indicators of cancer presence however richer contextual properties are only available at the intermediate level (i.e. ducts, lobules). In the case of tumour localisation, abnormal structures are more evident by analysing the relationship between cellular structures. Beck *et al.* [18] revealed features extracted from "tumour nests" were more informative than features extracted from individual cancer cells, highlighting the importance of contextual information in histopathology.

3.3 Tumour grading

The process of grading histological slides refers to the analysis of pathological properties of abnormal tissue to categorise cancer development. It provides an efficient method for summarising cancer risks through widely used scoring systems such as Gleason [54] or Nottingham [44]. Automation of tumour grading is a large and active research area which has developed considerably over the years.

The most straightforward approach to automatically categorising tissue into grades is to treat it as a classification problem using low-level textural features [42, 149]. However this approach is unsuitable in histopathology as textural properties are highly variable between histological samples and tissue structures. In a method proposed by Weyn *et al.* [149], a preprocessing step was required to limit classification to segmented cells thereby removing complexities involved with remaining tissue.

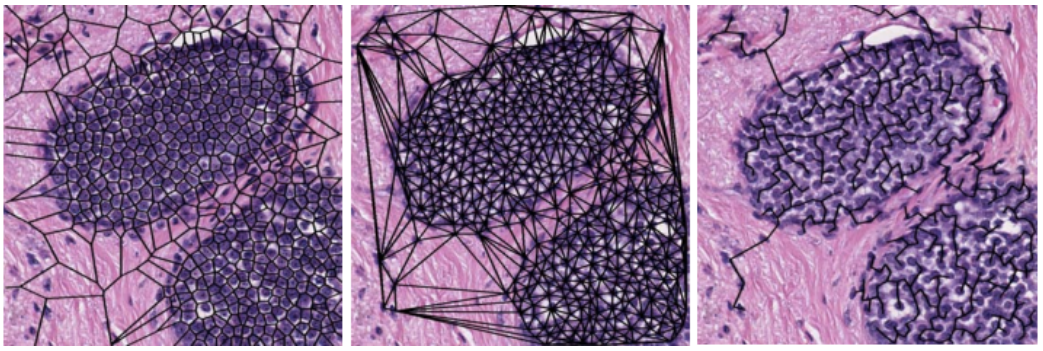


FIGURE 3.2: Voronoi (left), Delaunay (middle) and Minimum Spanning Tree (right) graphs overlaid on H&E stained low grade cancer tissue. Used with permission from Scott Doyle [39], © 2008 IEEE.

Graph models such as Voronoi diagrams, Deluanay Triangulation and Minimum Spanning Trees are the most common method for estimating tumour grades. The intuition is the relationship between segmented cell nuclei enable modelling of tissue components in a graph structure. Some techniques [15, 40, 102] combine cell segmentation methods (Section 3.2) with graph-based features to quantify relationships between cell nuclei. An alternative approach, avoiding cell segmentation, is to apply graph models directly to the tissue [6, 39, 41]. In doing so, all tissue components are captured in the final feature representation. In related work by Doyle *et al.* [39], graph and texture (grey-level, Haralick, Gabor) features were combined to grade breast tissue. Figure 3.2 shows how tissue structure was captured within graphs, with clustering of nodes inside ducts containing cancer and fewer nodes in stromal regions. However relationships between nodes in a graph structure can become complex with associated high computational costs, particularly in high-resolution histology images.

In other graph-based methods, Doyle *et al.* [40] described a pairwise classification approach for Gleason grading of prostate cancer. Grades were assigned through a refinement process whereby regions were iteratively classified into two groups starting with cancer vs. non-cancer, to grade 3 and 4 vs. grade 5 (cancer) and epithelial and atrophy vs. stroma (non-cancer), and so on. Here, the authors took advantage of grade groups which shared similar appearances. Alternatively, Basavanhally *et al.* [16] proposed a method for tumour grading in whole mount slides whereby features were extracted from multiple fields of views, thereby capturing multiscale information.

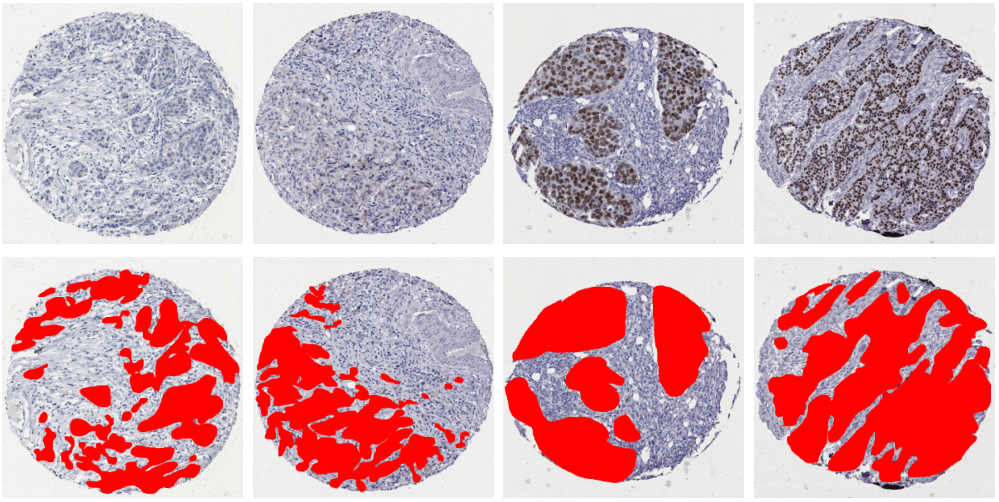


FIGURE 3.3: TMA spot images (top) and tumour regions annotated by an expert pathologist shown in red (bottom).

As TMAs are still at a stage of early usage in clinical research, there are few examples in the literature exploring grading of TMAs. Recently, Ali *et al.* [4] proposed a model for grading TMA prostate cancer images using a hybrid active contour model. Cellular structures were segmented by initialising an active contour via a watershed algorithm. As prostate cancer has a relatively rigid structure, active contours can appropriately trace the outline of tumour regions in these images. In other types of tissue such as breast and for invasive cancers, tissue structure and tumour boundaries are more complex.

3.4 Tumour localisation

In this thesis, tumour localisation refers to the identification of regions which encase cancerous cells. Tumour localisation is important in IHC assessment to measure the presence of specific biomarkers within cancerous tissue. Tumours which react to specific antigens can be treated accordingly, to reduce cancer development. Examples of tumour localisation in the form of hand-drawn annotations are shown in Figure 3.3.

Current methods for automated IHC assessment rely upon manual intervention to locate tumour in digital slides. For example, in Oncomark [108], the IHC-Mark Nuclear

algorithm requires manual intervention to highlight tumours for analysis. Specifically, a pathologist manoeuvres a pen tool at multiple magnifications to mark the outline of tumour regions. A similar markup technique is adopted in the Aperio IHC Nuclear algorithm [11]. The localisation of tumour is a challenging problem as there are currently no guidelines for detecting tumours, and cancer appearances vary considerably between types and grades. In this section, the literature on tumour segmentation and classification is reviewed.

Commercially, there are few viable options for automated tumour localisation. In TissueMark by PathXL, a tumour segmentation algorithm is provided for microdissection purposes [114]. As such, identified tumour regions are likely to encompass healthy structures, not suitable for IHC assessment. Furthermore, this feature is designed for whole mount slides and has not been applied to TMAs. The PathXL TMA toolbox [113] currently does not offer tumour segmentation for IHC assessment.

In the image processing research literature, there has been recent work in automatic tumour segmentation and classification. In almost all of these methods contextual information is captured in either the form of patches [72, 73, 104], multiple scales [26, 27, 42, 49] or reference locations [154]. K-means is a popular approach for clustering tissue types from RGB values for preprocessing [72, 73]. However, relying upon colour information in IHC stained samples can lead to errors when healthy epithelial cells exhibit positive staining. Instead, recently proposed methods extract low-level features from which codebooks are learned [56, 87, 104]. Learned codes provide richer representations of whole or partial tissue structures and are therefore suitable for locating tumour.

To capture wider context beyond a small region or pixel, Chang *et al.* [26] constructed spatial pyramids from sparse codes for image classification of whole mount slide images. Whilst this work has not been applied to tumour segmentation, results reported in [26] showed spatial pyramids capture large biological variations with few training samples. In TMAs, Xu *et al.* [154] used auto-context [137] in a Multiple Instance Learning (MIL) [38] setup to capture essential contextual information in patches. Here, cancer types were classified in images of human colon but the method is as

yet to be applied to breast TMAs. Foran *et al.* [49] captured pixel neighbourhoods in breast TMAs by constructing multi-scale texton histograms, where textons were derived from the Schmid filter bank [126] which also captures filter responses at multiple scales. In following chapters, Schmid filter banks are revisited (Section 5.3.2) and methods for capturing context at multiple scales are explored further (Section 6.3).

In other work exploring TMA analysis, Karaçali *et al.* [73] proposed a method for detecting regions of interest (ROIs) by modelling proportions of (RGB and k-means) segmented chromatin-rich, stromal and lumen regions. One of the limitations of this technique is the exclusion of textural and structural information in the tissue. Wang *et al.* [147] proposed a method which separated IHC stains from H prior to processing to reduce the impact of staining in tumour classification. However this method requires manual markup of regions for four tissue labels (tumour, stroma, lymphoid/inflammatory cells/necrosis, background) and this remains demanding in terms of staff resources needed.

Local image features have also been adopted in some methods [49], however performance tends to be poor and segmentations are noisy; texture alone fails to provide descriptive information of tumour appearance and surrounding tissue. More often low-level features are used as the basis for contextual representations. For example, Gorelick *et al.* [56] extracted RGB histograms in the form of annuli encasing superpixels. Khan *et al.* [75] proposed several high dimensional features extracted from tumour, hypocellular stromal and hypercellular stromal regions, which were then reduced using a modification of random projections [70].

For image-level classification, Li *et al.* [87] learned codebooks from low-level features in the form of randomised forest trees. Previously described work by Xu *et al.* [153] is also applicable at the image level. Both these methods are unsupervised i.e. expert opinion is not required during the learning process. Zhang *et al.* [157] proposed two-stage Support Vector Machine (SVM) and Multi-layer Perceptron (MLP) ensembles whereby "straight-forward" breast tissue images were classified using SVMs, leaving

more difficult cases for stage 2 or human experts (via a rejection option). This technique is reliant on an expert being available during both training and testing; given limited pathologist time, this may not be an option.

3.4.1 Tissue classification in digital pathology

In some digital pathology software, pattern recognition techniques have been adopted to classify different types of tissue. To some extent, this is also a form of tumour localisation as tumour labels can be assigned during training. In Leica Genie [84] and Indica Labs TMA software [66], a system can be trained to recognise various types of tissue (e.g. stroma, epithelial tissue, fat, tumour) by providing labelled data in the form of annotations. Once trained, the software classifies regions automatically which can then be used in image analysis algorithms. In Definiens Tissue Studio [35], tissue is segmented into "objects", after which each object can be assigned a tissue label. Similarly, objects are trained for classification of tissue types.

In practice, pathologists and technicians must commit considerable time to provide accurate and sufficient numbers of labels for pattern recognition software. In a study reported by Rizzardi *et al.* [123], altogether 11 hours of pathologist and technician time was required to train Aperio Genie. Notice that labels must be provided for various types of tissue in order for the software to be able to isolate tumour. Regular re-training is required for different laboratories, datasets and stains; this also requires additional acquisition of training images. Furthermore, when applied to TMAs, classification is poor due to high variability between samples and lack of training samples.

3.5 Other tumour image analysis models

Quantification of stromal cells has been an important indicator of invasive breast tumours. Beck *et al.* [18] found stromal morphometric features were a strong indicator for assessing patient survival, better than epithelial features in the case of grade 2 and 3

tumours. The authors suggest "tumour stromal" segmentation is important for extraction of prognostically informative features. Similarly, Linder *et al.* [91] differentiated stromal from epithelial cells for the purposes of IHC scoring.

In other work, mitotic cells have become an important indicator for prognosis of breast cancer. The Assessment of Mitosis Detection Algorithms (AMIDA) 2013 [142] challenge and MITOS 2013 [32] challenge revealed deep convolutional neural networks [50] performed favourably for mitosis detection compared to other methods which adopted hand-crafted features. However these models resulted in high computational costs and required considerable time to tune parameters without overfitting. More recently in the MITOS-ATYPIA 2014 challenge, the winning method [74] modelled mitotic cells within classified breast tumour regions using Gaussian mixture models [119]. As a preprocessing step, in [74] tumour regions were identified using random projections [75].

3.5.1 Gland segmentation

Another technique for analysing structural properties of breast, lung, colon and prostatic tissue is to segment glands and analyse them separately from surrounding tissue. For example, Nguyen *et al.* [106] segmented glands in prostate cancer to assign Gleason grades 3 and 4. One approach for gland segmentation is to adopt active contours at glandular boundaries [153]. Nguyen *et al.* [106] suggest the use of active contours is not suitable for gland segmentation as a gland does not have a fixed shape or size. As such, boundary approximations are difficult to acquire for initialisation of snakes. Instead, Nguyen *et al.* used the *Lab* colour space to identify lumen and epithelial nuclei which were enlarged and grouped to identify gland boundaries. Other methods exploit domain-specific knowledge about the structure of glands. In reported studies [99, 115], region-growing techniques were adopted where lumen in the centre of glands denoted seed locations. Alternatively, Sirinukunwattana *et al.* [130] classified small compact regions called superpixels to build glandular probability maps in colon images.

Gland segmentation is appropriate for *in-situ* cancers. However it is unsuitable for invasive cancers as glands are destroyed by cancer cells invading the surrounding tissue. The majority of techniques described above perform gland segmentation as a process of elimination of healthy glands as opposed to tumour detection.

3.6 Summary

From the literature reviewed, it is evident that tumour identification is still a challenging problem with various approaches adopted in previous work for detecting cancerous structures. For IHC assessment, it is essential to localise tumour regions to ensure accurate cell quantification. In commercial software, developers have overcome this issue by enabling manual intervention; a fully automated process can potentially provide many benefits. Automation introduces prospects of removing inter- and intra-observer variability, increasing throughput, and eliminating laborious manual tasks such as IHC scoring. Whilst explicit cell segmentation has its merits, classification of cells is poor due to lack of contextual information. Texture and appearance of tumour regions are more apparent at an intermediate level where multicellular structures can be captured. A summary of reviewed tumour segmentation methods which capture context using different techniques are outlined in Table 3.1.

TMAs are a relatively recent advancement in medicine; as such there has been little exploration in terms of automated analysis. However, recent work demonstrates potential. Difficulties in TMAs primarily arise due to highly variable samples and reduced numbers of training samples. In addition, annotations of TMAs are difficult to acquire at the pixel-level due to high resolution scans and limited availability of expert pathology input. Any automated solution must be able to operate on few training labels. Previous work in which multiple labels are required for various tissue types are unsuitable given pathologists' workloads.

Classification of breast cancers at present are limited perhaps due to the complex tissue structure. Some described methods model specific structures such as glands. However

Study	Dataset	Tissue	Method	Accuracy
Beck *et al.* [18]	1286 TMAs	Breast	Contextual/relational features extracted from superpixels; epithelial/stromal classification for patient survival	N/A
Chang *et al.* [26]	1380, 2148 images	Brain (GBM), kidney (KIRC)	Morphometric features extracted from segmented cell nuclei; image-level classification using spatial pyramids	92.91% (GBM), 98.50% (KIRC)
Foran *et al.* [49]	100 TMAs	Breast	Multi-scale texton histograms; Adaboost classification	$\sim$90%
Gorelick *et al.* [56]	50 WMS	Prostate	Colour, morphometric and SIFT features extracted from superpixel representation; trained using SVM	$\sim$85%
Karaçali *et al.* [73]	14 WMS	Breast	K-means clustering on greyscale and *Lab* color space; classification using estimated log-likelihood ratios	N/A
Khan *et al.* [75]	35 images	Breast	Extraction of Gabor features from stain normalised images; dimensionality reduction via random projection emsemble	F1 score: 0.89
Li *et al.* [87]	60 TMAs	Colon	Sparse codes constructed from partitioning of randomised trees	AUC: 0.987
Wang *et al.* [147]	9 TMAs	Lung	Texture features extracted from blue colour channel; Markov Random Fields applied to discrete labels to optimise tumour labelling	80%
Zhang *et al.* [157]	361 images	Breast	Texture features of which Curvelet Transform were superior; SVM ensemble for 2-class classification with rejection criteria	99.25%

TABLE 3.1: Overview of tumour segmentation/classification methods reviewed in Section 3.4. Details are given for the dataset used in each study (where WMS are whole mount slides and "images" are sub-images extracted from tissue slides), the type of tissue investigated, an outline of the method proposed and accuracy rates as reported in original papers.

this approach fails in high grade breast cancers whereby glandular structures are destroyed. A more general-purpose image analysis technique is required, applicable to cell nuclei, cytoplasm and membrane stains; as well as other tissue structures which have shown to be important for prognosis i.e. stromal cells.

To classify cells appropriately for IHC assessment, an analysis of the strength of the antibody stain is required. In histological samples where more than one stain is applied, stain normalisation can aid isolation of protein expression. Furthermore as tissue preparation differs between laboratories, stain normalisation also enables standardisation of stain intensities such that trained image analysis systems can be applied across multiple datasets. Whilst the author is aware of the literature in stain normalisation [76, 93, 94], it is not the focus of this thesis. Investigation of the usage of stain normalisation is reserved for future work (Section 9.4).

In this research, image analysis techniques are investigated for the purpose of localising tumour in images of TMAs, with the aim of performing IHC assessment. From previous work in tumour localisation, contextual information has been shown to successfully capture cancerous structures. As such, methods described in the following chapters model contextual information in a rotation invariant manner, suitable for histopathology. Before approaching the task of automated tumour localisation, manual localisation of tumours is first explored.

Chapter 4

Manual Tumour Localisation

4.1 Introduction

Inter-rater agreement in clinical medicine refers to the agreement between two or more specialists when performing tasks such as IHC assessment. By measuring the inter-observer agreement, we can determine how much clinical measurements concord when performed manually. This knowledge is essential for evaluating the current "gold standard" amongst pathologists in clinical practice. However, in the computer vision literature, inter-rater agreement is seldom reported. Often this is due to lack of resources, and specifically the resources required to acquire annotation labels from multiple sources which in the medical domain, can be expensive.

The current method for manual IHC assessment, specifically IHC scoring, relies upon an ordinal scale (Section 2.5.1). A study in 2007 [13] reported inter-rater agreements for ER and PR Allred scoring in 89 consecutive cases of invasive ductal carcinoma of the breast (Table 4.1). Whilst percentage of positive cells resulted in "substantial" agreement, agreements between IHC scores was at best "fair" which can have an adverse effect in research and patient care (i.e. treatment decisions). In IHC analysis, disagreements arise due to ambiguous scoring cases e.g. distinguishing "weak" and "moderate" staining strengths. In the case of proportion scores, percentages of positive

Scores	κ	
	ER	**PR**
Allred scoring		
Total	0.340	0.451
Proportion	0.478	0.596
Intensity	0.441	0.493
Percentage of positive cells	0.673	0.725

TABLE 4.1: Evaluation of Fleiss κ inter-rater agreement in 89 cases of invasive breast cancer. [13]

cells are usually roughly estimated without counting and are therefore highly subjective [140]. Despite these variations, manual analysis is necessary as expert knowledge is required to interpret biological materials. Furthermore, the scoring technique is relatively simple, low cost and enables measurements to be acquired quickly [13].

As described in the previous chapter, image analysis algorithms in digital pathology can potentially provide more accurate and standardised measures compared to manual analysis. To assess the current benchmark for manual localisation of tumour, a study was designed to measure the inter-rater agreement between hand-drawn annotations of tumours in TMAs. To the best of the author's knowledge, this is the first study which evaluates inter-rater agreement for the purpose of tumour localisation in histopathology images.

In this chapter, inter-rater agreement is reported between segmentation masks obtained from two expert pathologists. By measuring how much experts agree with each other, some insight can be gained into pixel-level accuracy required to maintain current manual IHC assessment. Results reported here will form a benchmark for automated tumour localisation methods reported in subsequent chapters. Furthermore, as hand-drawn segmentations are unlikely to be accurate to the pixel level, a novel technique to categorise disagreements between binary segmentation masks is also described.

4.2 Materials and Methods

4.2.1 Tissue microarray data

Four breast TMAs were generated from primary, previously untreated breast cancers held under the delegated authority of the Tayside Local Research Ethics Committee in the Tayside Tissue Bank, Dundee, UK. Briefly, surgically resected primary breast cancer from otherwise unselected patients was fixed in buffered formalin, paraffin-embedded and stored at a controlled temperature (18-22°C) overnight then further processed to formalin-fixed paraffin-embedded blocks. Whole mount sections stained with H&E were marked to highlight relevant invasive cancer or normal tissue to allow TMA generation of up to six 0.6mm cores per cancer. TMAs were then constructed using a manual tissue arrayer (Beecher Instruments Inc., Sun Prairie, WI, USA). Four-micron TMA sections were cut, mounted onto poly-L-lysine coated glass slides and subjected to nuclear staining for ER using a Novocastra antibody. Stained slides were scanned using an Aperio Scanscope XT (Aperio Technologies CA, USA) on a x20 objective with an optical doubler in place (equivalent to x40 optical objective). Each slide was then segmented into the individual constituent stained spots; each spot represents a section from a tissue core.

Thirty-two uncompressed TIFF format images of TMA spots from thirty-two breast cancers were used. The perimeter of each spot was delineated and pixels exterior to this perimeter were excluded from subsequent analyses. Each spot image was approximately 3000 pixels in diameter and contained invasive tumour regions.

4.2.2 Manual segmentation of tumour regions

Tumour regions in the TMA spots were manually segmented using Aperio Technologies' Spectrum Software with TMA Lab and the webscope interface (Aperio Technologies, CA, USA). This task was performed on a Wacom Bamboo Fun tablet (model CTH-461) using a stylus for precision. Segmentation involved manually tracing the

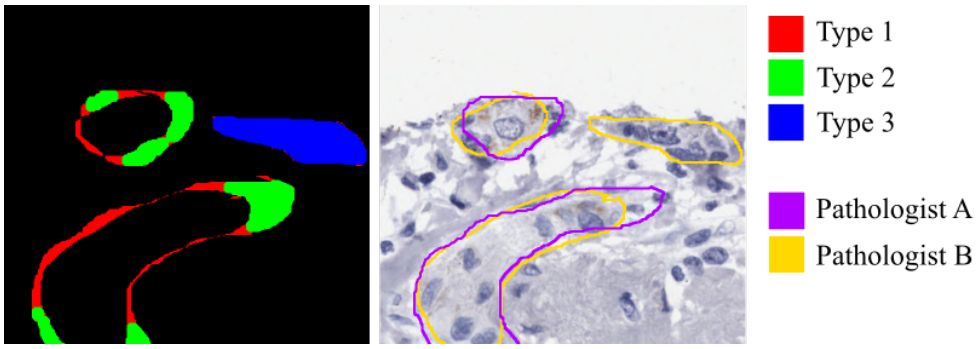


FIGURE 4.1: Examples of Type 1 (red), Type 2 (green) and Type 3 (blue) disagreements. Annotations drawn by pathologist A (purple) and pathologist B (orange) are shown on the right overlaid on the original image.

boundaries of invasive tumour regions; the software tool displayed filled regions overlaid on the TMA spot images as they were annotated. Pathologists were instructed to interact with Aperio as they would normally to avoid altering the manner with which "typical" annotations are acquired.

Each spot was annotated independently by two qualified specialist pathologists, Dr. Lee B. Jordan (LBJ) and Dr. Colin A. Purdie (CAP), resulting in two sets of tumour masks. Each mask labels each pixel as either tumour (T) or non-tumour (N). In the remainder of this thesis, LBJ will be referred to as pathologist A, and CAP as pathologist B. Both pathologists have several years of expertise in the field of histopathology and are currently appointed in the Department of Pathology, Ninewells Hospital, U.K.

4.2.3 Comparing spot segmentations

Each spots' manual segmentation masks were compared with each other by comparing pixel-level labels in each of the TMAs. However, there are qualitative differences between segmented regions that are not well captured by pixel-level analysis. Therefore, when comparing two segmentation masks, pixels were categorised into three types of disagreement: Type 1, Type 2 and Type 3. Illustrations of these disagreements are shown in Figure 4.1.

To compare two tumour segmentation masks, S_A and S_B, two binary difference images, D_{A-B} and D_{B-A}, were produced. 8-connected morphological opening was then applied to D_{A-B} and D_{B-A}, using a circular structuring element with a radius of 10 pixels, resulting in O_{A-B} and O_{B-A}, respectively. The structuring element approximates the size of a small epithelial cell. Type 1, Type 2 and Type 3 disagreements are described as follows.

Type 1 Region boundaries in two segmentation masks are often separated along part of their lengths by distances of only a few pixels. Such discrepancies may arise from a lack of precision when using the stylus and/or from the lack of any clear visual boundary to annotate in the image. As such they are likely to be inconsequential for subsequent tasks such as IHC scoring because such small separations do not allow for the inclusion or exclusion of entire cells.

Definition: Type 1 disagreements are pixels removed during the opening process i.e. pixels in D_{A-B} which did not appear in O_{A-B}. Similarly, the same comparison was performed between D_{B-A} and O_{B-A}. Pixels that differed before and after the opening operation were labelled Type 1.

Type 2 Disagreements which are not of Type 1 are large enough to encompass epithelial cells (Figure 4.1). A pixel disagreement is labelled Type 2, if it is not of Type 1 and it is in a region labelled as tumour in one mask which overlaps with a region labelled as tumour in the other mask. Type 2 disagreements can arise from differences of opinion about the spatial extent of a tumour region.

Definition: Region(s) in O_{A-B} and O_{B-A}, identified via 8-connected component analysis, which connect with agreed upon tumour region(s) were labelled Type 2.

Type 3 Disagreements that are neither Type 1 nor Type 2 are designated Type 3, reflecting differences of opinion about whether or not a group of cells is malignant.

Disagreement types were visualised by computing difference images from pairs of segmentation masks and then colour-coding pixels for which the segmentations differed as Type 1 (red), Type 2 (green) or Type 3 (blue).

		Pathologist A	
		T	N
Pathologist B	T	**0.270**	0.049
	N	0.043	**0.638**

TABLE 4.2: Normalised contingency table comparing segmentation labels in masks produced by pathologist A and pathologist B.

Only pixels within each TMA spot were analysed; background pixels were ignored. Background pixels were identified manually in the form of a spot background mask.

4.2.4 IHC scoring

To evaluate the impact of using manually obtained segmentation masks for IHC assessment, IHC scores were computed using the FDA-approved Aperio IHC Nuclear Version 10 algorithm (Aperio Technologies, CA, USA). Only regions labelled as tumour were passed to the scoring algorithm. The Aperio IHC algorithm identifies nuclei automatically and outputs a staining intensity score (ranging from 0 to 3) and an estimate of the percentage of positively stained cells. From these measurements, IHC (Allred and Quickscore) scores were computed.

4.3 Results

A comparison of the segmentation masks produced manually by pathologists A and B is summarised in Table 4.2 in the form of a normalised contingency table. The inter-rater agreement between the two pathologists was $\kappa = 0.908$. Proportions of false positive and false negative disagreements were close to equal ($4-5\%$).

A set of disagreement visualisations comparing annotations from both pathologists is shown in Figure 4.2. Image patches showing disagreements in more detail are shown in Figure 4.3. Visual assessment shows disagreements varied considerably between TMA spots. In Figure 4.2(d) and Figure 4.2(e), a large proportion of Type 3 disagreements are visible. However in Figure 4.2(a), no Type 3 disagreements are present but

Type 1	Type 2	Type 3
0.227 ($\pm$0.144)	0.593 ($\pm$0.218)	0.180 ($\pm$0.227)

TABLE 4.3: Proportion of Type 1, Type 2 and Type 3 disagreements between segmentation masks obtained from pathologist A and pathologist B. Standard deviations are given in brackets.

Scores	Allred	Quickscore
Total	0.843	0.885
Proportion	0.877	0.921
Intensity	0.944	

TABLE 4.4: Evaluation of inter-rater Fleiss κ agreement between intensity, proportion and total (i.e. sum of intensity and proportion) Allred scores and Quickscores.

a large proportion of Type 2 disagreements are. A qualitative comparison between pathologists' annotations showed more refined boundaries in pathologist A's segmentation masks, whereby more pixels were excluded from tumour regions. Figure 4.2(a) shows one case when this difference was most noticeable. However, in the majority of segmentation masks, drawn tumour regions were similar between pathologists. Figure 4.2(b) and Figure 4.2(c) show examples of annotations in which few disagreements between pathologists were identified; here the majority of disagreements correspond to Type 1 disagreements.

Table 4.3 summarises the distribution of pixels corresponding to Type 1, Type 2 and Type 3 disagreements over TMA spots in the dataset. Over 20% of disagreements corresponded to minor misalignment along tumour boundaries. These disagreements are unlikely to effect IHC scores computed from segmentation masks. Over 59% of disagreements were of Type 2, corresponding to regions where cells can be present. The proportion of Type 3 disagreements was considerably lower. However, note the high standard deviations, which confirm the high variability of disagreement types between TMA spots.

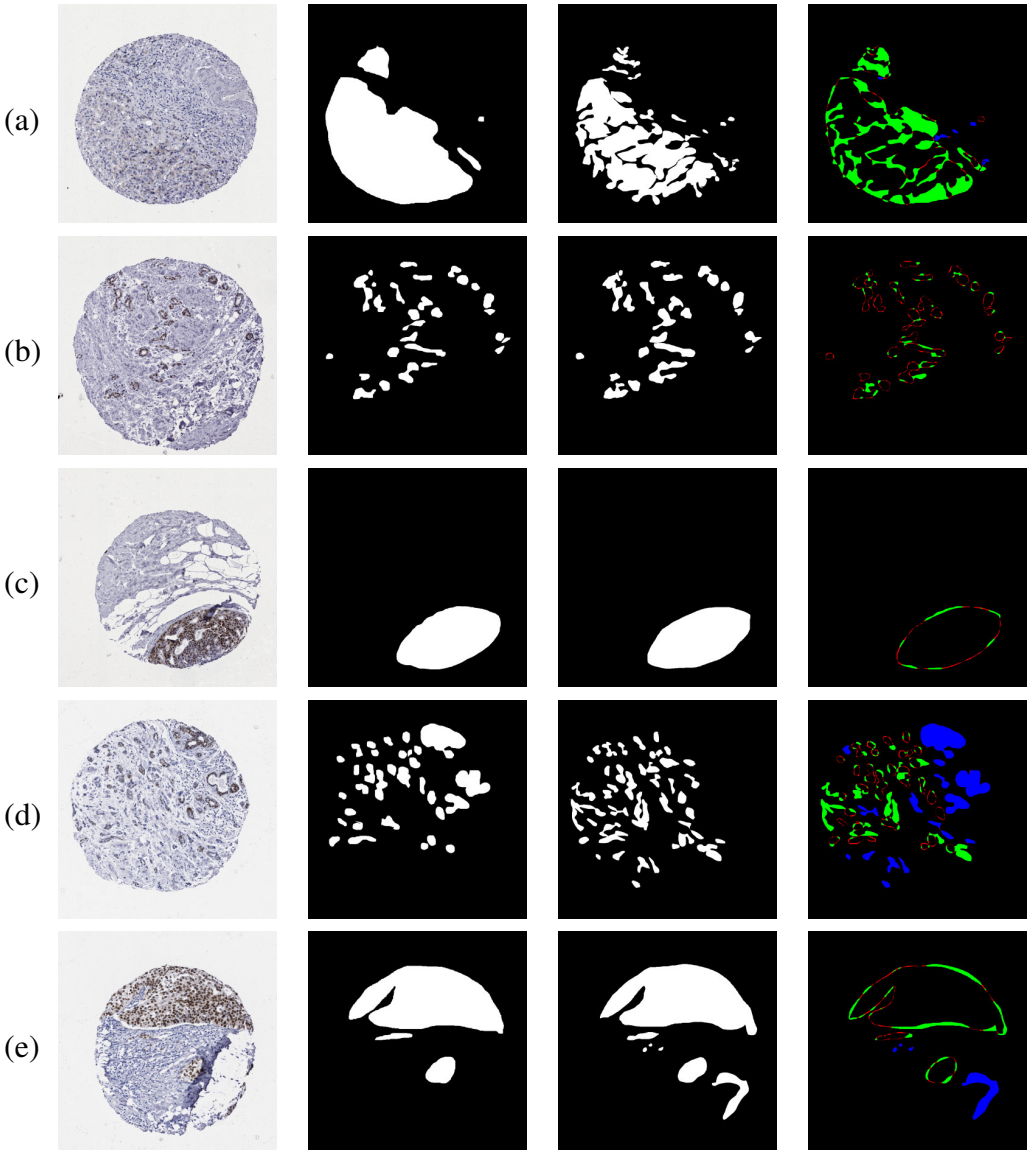


FIGURE 4.2: TMA spot image (left), manual segmentation masks from two trained pathologists (pathologist A: centre right, pathologist B: centre left) and colour coded difference image (right). Regions are labelled according to Type 1 (red), Type 2 (green) and Type 3 (blue) disagreements.

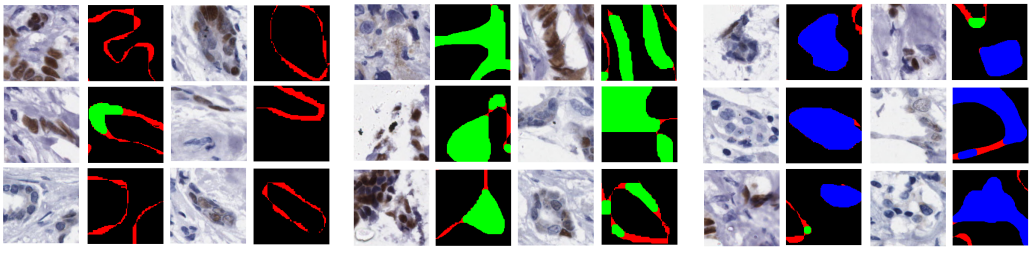


FIGURE 4.3: Image patches (left) and corresponding Type 1, Type 2 and Type 3 disagreements (right) between pathologists.

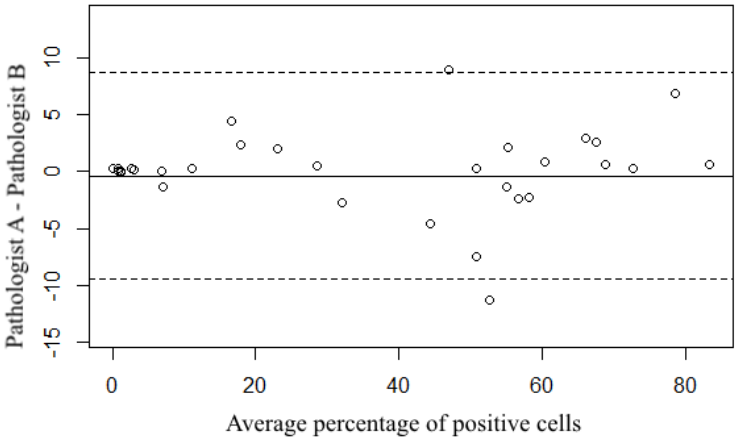


FIGURE 4.4: Bland Altman plot of percentage of positive cells identified in the Aperio software.

4.3.1 IHC scoring

Inter-rater agreements for Allred scores and Quickscores are shown in Table 4.4. The ordinal scales for computed intensity scores in Allred and Quickscore scoring systems are identical (Section 2.5.1). Agreement between intensity scores were strong. Agreement for proportion scores were slightly higher for Quickscore compared to Allred. When comparing scoring systems, Quickscore allocated six scales for proportion whereas Allred allocated five (Section 2.5.1). Where proportion scores differed between pathologists, percentage of positive cells were on the boundary of proportion score ranges; as such proportion scores differed by at most one. A Bland Altman plot of percentages of positive cells is shown in Figure 4.4 where the standard deviation was $\pm 9\%$. For the majority of TMA spots, there were strong agreements between percentage of positive cells with differences close to zero.

The following experiment was designed to test the hypothesis that Type 1 disagreements are unlikely to effect IHC scores as a cell cannot fit within these regions. Type 1 disagreements were discounted from the segmentation masks (i.e. labelled as non-tumour and therefore not passed to Aperio). IHC scores were then re-computed from updated segmentation masks and compared to IHC scores retrieved from original segmentation masks. κ agreements between scores computed with and without Type 1

Scores	Allred	Quickscore
Total	0.924	0.962
Proportion	0.960	1.000
Intensity	0.944	

TABLE 4.5: Fleiss κ agreement between pathologist A's segmentation masks with and without Type 1 disagreements.

Scores	Allred	Quickscore
Total	0.959	1.000
Proportion	0.958	1.000
Intensity	1.000	

TABLE 4.6: Fleiss κ agreement between pathologist B's segmentation masks with and without Type 1 disagreements.

disagreements are shown in Table 4.5 (pathologist A) and Table 4.6 (pathologist B). When Type 1 disagreements were removed, computed Quickscores revealed slightly higher inter-rater agreements compared to Allred scores, particularly in the case of pathologist B's segmentation masks. Table 4.6 shows Quickscores computed from pathologist B's segmentation masks were unaffected across all TMA spots. Only one Allred score differed by one, resulting in $\kappa = 0.959$. In the case of pathologist A's segmentation masks, Quickscore proportion scores were identical and intensity scores resulted in high inter-rater agreements $(\kappa = 0.962)$. Overall, removal of Type 1 disagreements resulted in identical or similar IHC scores.

4.4 Summary

In this chapter, a study was described whereby manually hand-drawn tumour segmentation were compared between two expert pathologists with the aim of measuring the current benchmark of tumour localisation amongst pathologists. It was shown that there was a strong agreement between the two experts, resulting in pixel-level agreement of $\kappa = 0.908$ (Table 4.2).

To measure the impact of tumour localisation on IHC scoring, an experiment was designed to compute IHC scores from manual segmentation masks. Results showed inter-rater agreements (Allred: $\kappa = 0.843$, Quickscore: $\kappa = 0.855$), differed by at most one score. Minor misalignment between manual annotations, termed Type 1 disagreements, had little impact on extracted IHC scores.

Note that conclusions drawn in this chapter reflect segmentation masks generated by trained specialists employed for the study. It is anticipated agreements between laboratories and annotators may differ. Future work will investigate other factors which can impact manual IHC scoring.

In subsequent chapters the inter-rater agreement reported in this chapter will be used as the benchmark performance in the reported dataset. In the following chapters, methods to automatically localise tumour are explored. The inter-rater agreements described in this chapter are revisited to compare automated and manual segmentation masks (Chapter 7).

Chapter 5

An Extension and Evaluation of Spin-Context

5.1 Introduction

In this chapter, a technique called spin-context is described which extends the auto-context method described by Tu and Bai [137]. Here, "context" refers to the posterior distribution in a local neighbourhood fused with low-level appearance features. To make this method applicable for histopathology image analysis, context locations correlate to points on circular rings. In spin-context, the problem of locating breast cancers in images of TMA spots is formulated as classifying each location on a regular grid as being tumour (T) or non-tumour (N).

This work was done in collaboration with Dr. Telmo Amaral *(Culture Lab, Newcastle University, UK)*. Contributions in this thesis are (a) the extension of spin-context to incorporate TMA spots boundaries (Section 5.5), and (b) evaluation of spin-context, including a comparison between spin-context and auto-context (Section 5.7).

Before describing spin-context, related work in auto-context is reviewed in Section 5.2. Local image feature adopted in reported experiments are described in Section 5.3. The original implementation of auto-context is described in Section 5.4.1, followed by

the spin-context adaptation (Section 5.4.2). Spin-context is then extended to remove background interference (Section 5.5) and a comprehensive evaluation is performed (Section 5.7).

5.2 Related work

In computer vision, contextual information has been captured in many forms including Markov Random Fields (MRFs) and conditional MRFs (CRFs). These models operate on limited function families in which energy functions are maximised. Alternatively, shape context [19] captures context in a local neighbourhood from low-level feature descriptors. However prior information of the shape to be recognised is required. Whilst shape context can also be learned [69] or approximated [148], auto-context offers a simple approach to capturing context which seamlessly integrates into the classification procedure. Context features are extracted from posterior distributions directly from classification maps, thereby keeping computational costs low.

The auto-context framework described by Tu and Bai [136] extracts and concatenates posterior probabilities from key context locations resulting in a one-dimensional context descriptor. Context locations which contribute towards the context descriptor are selected using a star-shaped stencil. A detailed description of auto-context is provided later in this chapter (Section 5.4.1).

Auto-context has been applied to various applications since its origination in 2008 [136]. Monoz *et al.* [101] proposed a "stacked hierarchical scene labelling" method which iteratively partitions an image into refined superpixel parts. Posterior probabilities from superpixels and neighbouring superpixels in parent regions represent context. In earlier work by Poole [118], pre-dating auto-context, posterior probabilities were used in a similar manner for gathering context using a "9x9 square stencil". Here, distributions of classes were estimated from local neighbourhoods and modelled as a probability tree. More recently, Jampani *et al.* [68] used a stack of decision tree classifiers, the input of which consisted of image features and auto-context statistics.

To avoid overfitting, a stacked generalisation technique was adopted such that auto-context statistics were computed using different classifiers from the test fold.

In follow-up work by Tu and Bai [137], auto-context was evaluated on 3D brain MRI imaging, showing the potential of this technique in the medical domain. Since then, it has been applied in various medical applications. In whole-body CT, Montillo *et al.* [100] proposed an extension of decision forest classifiers that incorporated semantic context in a manner similar to auto-context. Li *et al.* [88, 90] combined context from regularly acquired treatment images (i.e. CT scans) with a planning image for prostate segmentation. Only probabilities with high confidences were updated and selected for context locations. Furthermore, Jurrus *et al.* [71] detected neuron membranes in electron microscopy directly from image patches (i.e. without feature extraction) in the auto-context framework. None of the above used distribution-based context descriptors and, appropriately for the applications considered, descriptors were not invariant under image rotation.

Auto-context has also been applied to 2D histopathology images to improve classification of class labels in tissue. Chomphuwiset *et al.* [30] used Hough transform-based techniques to detect cell nuclei in liver histopathology images. They also integrated random forest classification results, obtained from texture features, with context information from nearby nuclei and regions. Xu *et al.* [154] proposed a tumour segmentation, clustering and classification method using Multiple Instance Learning (MIL) for colon histopathology images. Contextual information was introduced as a prior for MIL to encourage neighbouring image patches to share similar class labels. However, to the best of the authors' knowledge, context has not been applied to breast histopathology images which is "unanimously considered a highly heterogeneous disease" [143]. This introduces difficulties when classifying small areas of breast tissue such as TMAs.

5.3 Image features

Low-level local image features, specifically spin intensity features [82] and differential invariants [126], were used in the following implementation of spin-context. Features were selected to ensure rotation invariance, however spin-context is not limited to these techniques. Other low-level features such as SIFT or HoG can also be integrated with ease, making spin-context adaptable to other classification problems.

5.3.1 Spin intensity features

Spin intensity image features were proposed for texture representation by Lazebnik *et al.* [82]. A spin feature encodes the distribution of brightness values within a circular support region centred at a location, h_0. Here, pixels within the support region (stored in a vector $\mathbf{h}$) are indexed by u. The spin feature is encoded in a rotation invariant histogram representation with two dimensions: the distance between each pixel $\mathbf{h}_u$ and h_0, $||\mathbf{h}_u - h_0||$, and the intensity value of $\mathbf{h}_u$, $I(\mathbf{h}_u)$.

As the spin histogram is a "soft histogram", each pixel contributes to more than one bin. The contribution of a pixel $\mathbf{h}_u$ to bin (d,i) is shown in (5.1). α and β are parameters that determine bin sizes, where each bin is indexed by the radial distance interval, d, and intensity interval, i. c_d and c_i denote the centre of corresponding distance and intensity bins, respectively. The resulting spin histogram, H, is a summation over u, $H_{di} = \sum_u w_{di}(h_u)$, where

$$w_{di}(h_u) = \exp\left(-\frac{(||\mathbf{h}_u - h_0|| - c_d)^2}{2\alpha^2} - \frac{(I(\mathbf{h}_u) - c_i)^2}{2\beta^2}\right) \tag{5.1}$$

5.3.2 Differential invariants

Differential invariants were computed by convolving image patches with a set of first- and second-order 2D Gaussian derivative kernels at three scales, using a Gaussian

pyramid [22]. Resulting convolutions were then combined to obtain four differential invariants at each pixel location [126]. The vector of differential invariants, $\mathbf{v}_i$, for location i, is defined in (5.2).

$$\mathbf{v}_i = \begin{bmatrix} L \\ L_x L_x + L_y L_y \\ L_{xx} L_x L_x + 2 L_{xy} L_x L_y + L_{yy} L_y L_y \\ L_{xx} + L_{yy} \\ L_{xx} L_{xx} + 2 L_{xy} L_{yx} + L_{yy} L_{yy} \end{bmatrix} \tag{5.2}$$

L is the luminance function convolved with a Gaussian, and the indices x and y represent the derivative with respect to the variables x and y, respectively. The first element of $\mathbf{v}_i$ is the zeroth order term.

In reported experiments, Gaussian derivative kernels had standard deviations of 8 pixels, and thus effectively 16 and 32 pixels and the second and third scales, respectively. Standard deviations were selected to incorporate parts of nuclei, whole nuclei and immediate surroundings. This setup is as described by Amaral *et al.* in [8].

5.4 Relevant context-based descriptors

5.4.1 Auto-context

Auto-context, described by Tu and Bai [137], is an iterative pixel labelling technique, in which label probabilities at a given iteration are used as contextual data for the following iteration. Contextual data are concatenated with local image features to form input vectors for each iteration. Context locations are chosen by applying a star-shaped "stencil" to labelled probability maps. An illustration of this framework is shown in Figure 5.1 and pseudocode is provided in Algorithm 1.

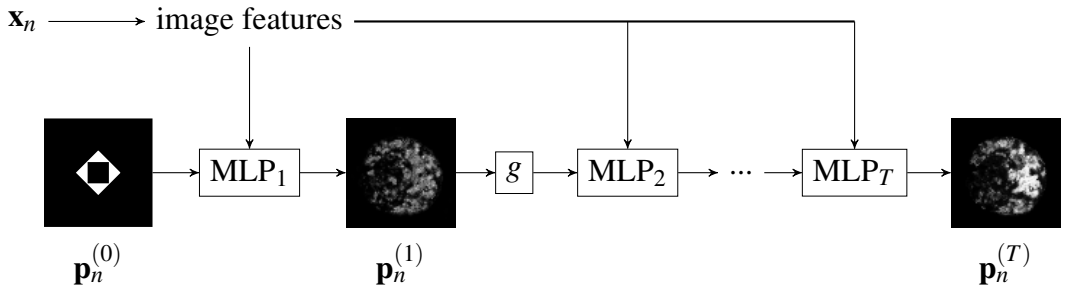


FIGURE 5.1: Auto-context classification of an image, $\mathbf{x}_n$ for T iterations. In each iteration an updated classification map, $\mathbf{p}_n^{(t)}$, is produced from classification of context descriptors and image features.

Algorithm 1 Auto-context training.

Given a set of N training images, $\mathbf{x}_1 \ldots \mathbf{x}_N$, together with their label maps, $S = \{(\mathbf{x}_n, \mathbf{y}_n), n = 1...N\}$:

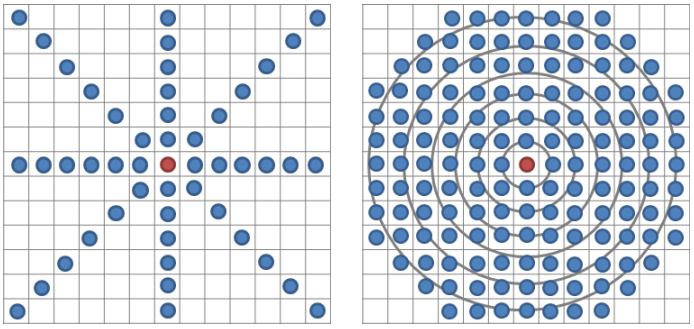
For each image $\mathbf{x}_n$, construct a probability map $\mathbf{p}_n^{(0)}$ containing M grid locations, with uniform distribution on all the labels.

For iteration $t = 1 \ldots T$:

1. Make a training set $S_t = \left\{(\mathbf{y}_{mn}, \mathbf{f}_{mn}, g(\mathbf{p}_n^{(t-1)}, m)), m = 1 \ldots M, n = 1 \ldots N\right\}$ where $\mathbf{f}_{mn}$ is the feature representation for $\mathbf{x}_{mn}$ and $g(\mathbf{p}_n^{(t-1)}, m)$ is the context descriptor at iteration $t-1$.
2. Train a classifier on S_t.
3. Use the trained classifier to compute new classification maps $\mathbf{p}_n^{(t)}$ for each training image $\mathbf{x}_n$.

In Algorithm 1, M is the total number of grid locations in the probability map $\mathbf{p}_n$. g is a function which computes a context descriptor from posterior probability values by selecting locations centred around location m. In iteration t, posterior probability values are selected from classification map $\mathbf{p}_n^{(t-1)}$. Figure 5.2(a) shows how context locations are selected using a star-shaped stencil. The red grid point denotes location m and blue locations are those at which posterior probabilities contribute to the context descriptor.

The prior, $\mathbf{p}_n^{(0)}$, is a uniform distribution. Both local image features and probability values are input into the classifier for training, which is subsequently used to output



(a) (b)

FIGURE 5.2: (a) Star-shaped stencil and (b) circular stencil for selecting context locations from label probability maps.

an updated classification map for iteration t. The algorithm iteratively updates T times, producing a series of T classifiers.

5.4.2 Spin-context

Tu and Bai [137] used a star-shaped stencil (Figure 5.2(a)) to select context location points around the pixel being classified. The resulting context features from this stencil were not invariant under image rotation. Spin-context [7] is an extension to auto-context which extracts context in a rotation invariant manner. In spin-context, context features for a given grid location are computed from label probability values within a circular support region. Spin-context is extracted analogously to intensity spin features, computing a two-dimensional soft histogram reflecting the distribution of probabilities within the support region, with rows representing probability intervals and columns representing radial distance intervals. Figure 5.2(b) shows the circular mask used to compute spin-context. Each ring corresponds to a radial distance interval in the resulting spin-context descriptor.

In iteration 1, context is not available from the previous iteration so a uniform constant descriptor is adopted. Therefore, $\mathbf{p}_n^{(1)}$ does not incorporate context. In subsequent iterations, context features in the form of a soft histogram are computed from $\mathbf{p}_n^{(t-1)}$.

As in auto-context, context and local features are concatenated in each iteration, t, to produce an updated classification map, $\mathbf{p}_n^{(t)}$.

Compared to auto-context, spin-context does not capture spatial clustering of similar structures e.g. clusters of cancer cells in ducts are not distinguishable from scattering of cancer cells in a larger region. Spatial information captured in the spin-context context descriptor is relative to a single point i.e. the centre of the stencil. However by not enforcing specific context points as in auto-context, spin-context offers additional benefits, described as follows.

5.5 Boundary sensitive spin-context

The spin-context descriptor has a desirable property, whereby context outside the tissue spot's boundary can be disregarded whilst only considering context within the spot region. Figure 5.3 illustrates the advantage of using spin-context to produce a more accurate representation of context information around the boundaries of the spot. Only blue context locations contribute towards the context histogram; orange points are ignored.

The use of boundary information prior to context extraction allows contributions of context points outwith the TMA spot to be ignored, when constructing the normalised two-dimensional spin histogram. In doing so, not only is context information accurate for the current iteration but subsequent iterations also reflect accurate information extracted from the spot region. The star-shaped stencil context descriptor, not being distribution-based, does not allow this level of flexibility to be maintained, resulting in background interference, or conversely the need to handle missing context data.

TMA spot regions take the form of binary segmentations where 1 denotes tissue and 0 are background pixels. During context extraction, each grid in the binary segmentation centred on the mth location is compared with the circular stencil to identify relevant grid locations. Resulting context descriptors near the spot boundary therefore reflect only locations containing tissue.

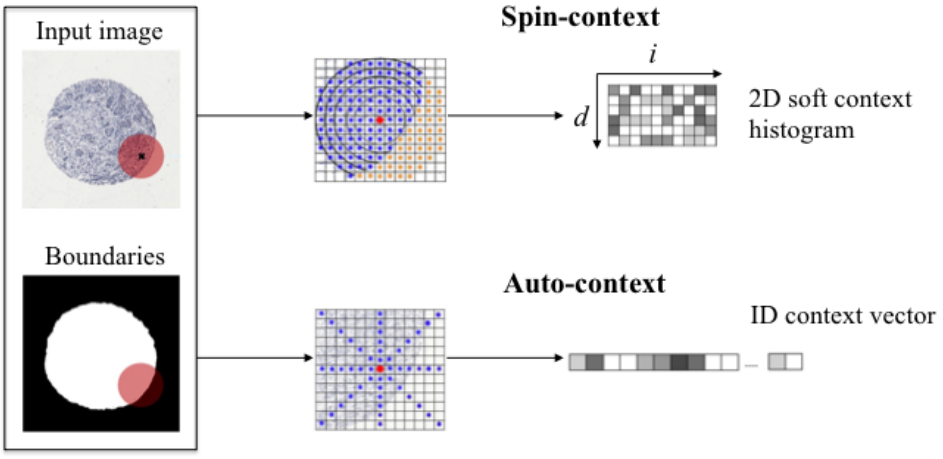


FIGURE 5.3: A binary mask is used to ignore the contributions of pixels outside the spot's boundary to the spin histogram. Auto-context corresponds to label probability values at all locations lying on a star-shaped stencil, regardless of spot boundaries.

5.6 Experiments

32 TMA spots containing tumour regions were subjected to nuclear staining for ER. Spot images were 3600 x 3600 pixels. Manual annotations were retrieved as described in Chapter 4. Manual labels from pathologist A were used in the following experiments, however a comparison between two expert pathologists (pathologist A and pathologist B) is provided in Section 5.7.1.

Tumour labelling was evaluated using 8-fold cross-validation on the 32 spots. Each cross-validation experiment was repeated eight times to measure variability. Multilayer percepton (MLP) classifiers were used with five hidden units, a regularisation constant of 0.1 and scaled conjugate gradient optimisation. MLPs were trained to output class posterior probabilities. Local and context features were computed at points on a 136 x 136 grid (a grid step of 25 pixels). Differential invariant features were computed at three scales using a Gaussian pyramid and filters with a standard deviation of 8 pixels. Intensity spin local features were computed at two scales with a circular support region with a radius of 50 pixels. Spin-context used a circular support region with a radius of 6 grid points, as shown in Figure 5.2(b). To evaluate boundary-sensitive spin-context, hand-drawn TMA spot segmentations were generated.

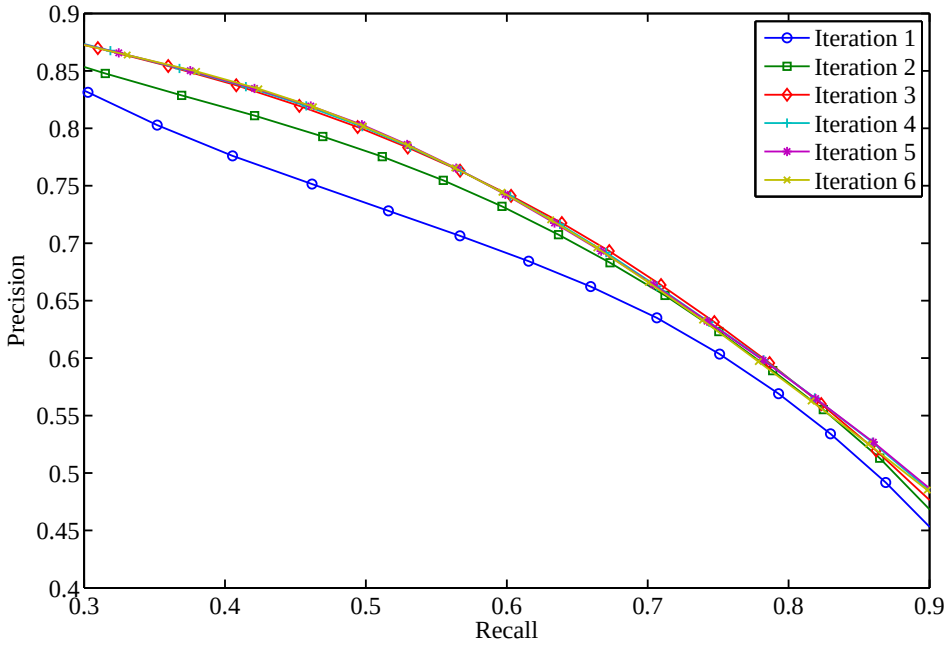


FIGURE 5.4: Precision-recall curves for six spin-context iterations.

Auto-context (non-rotationally invariant context) was also evaluated using a stencil in which neighbouring grid points within a radius of 6 grid spacings in each of the 8 cardinal and inter-cardinal compass directions were used as context, as shown in Figure 5.2(a). The auto-context stencil shares the same context window size as the spin-context stencil; as such both these methods are comparable in reported experiments.

5.7 Results

The precision-recall curves in Figure 5.4 displays the results obtained for six spin-context iterations. In the first three iterations, a noticeable improvement is shown; curves converge in subsequent iterations, suggesting three iterations are sufficient to incorporate contextual information. After this, there is little performance gain. Compared to a standard MLP classifier which incorporated no context (i.e. iteration 1), spin-context improved the precision-recall curve.

	AUC	95% CI
Spin-context	**0.926**	(0.924, 0.927)
Auto-context	0.923	(0.920, 0.925)
No context	0.916	(0.914, 0.918)

TABLE 5.1: AUC values for no context and three iterations of spin-context and auto-context. AUC values are reported as a mean between eight repeated experiments. Upper and lower confidence intervals (CI) are also reported.

Figure 5.5 shows a subset of results achieved using spin-context. Figure 5.5(a)-5.5(d) shows successful tumour labelling in which the introduction of spin-context improved classification performance. In Figure 5.5(a), lower tumour probabilities when no context was incorporated, were (correctly) significantly higher after six iterations. Figure 5.5(e) and 5.5(f) show unsuccessful labelling. In these cases, initial classification maps were poor which was reflected in subsequent spin-context iterations. In another auto-context technique described by Jampani *et al.* [68], similar outcomes were observed between various pixel predictions. The authors argued good pixel prediction is important in an auto-context framework, even as an intermediate step.

Figure 5.6 compares spin-context with stencil-based auto-context. Standard deviation bars, generated from eight repeated experiments (each with eight folds), are shown as dotted lines. Spin-context and auto-context showed similar performance, with spin-context excelling slightly at lower recall values. Both methods also showed an improvement compared to a classifier with no context. AUC values for the same experiments are also reported in Table 5.1. At first glance, mean AUC values suggest spin-context surpassed auto-context but confidence intervals suggest performance can, on occasion, be similar.

5.7.1 Comparison with manual annotations

In Chapter 4, an inter-rater agreement of $\kappa = 0.908$ was reported between expert breast pathologists, pathologist A and pathologist B. Any automated solution must be able to show similar or higher agreements to potentially replace manual input. Table 5.2 and Table 5.3 show normalised contingency tables for three iterations of spin-context

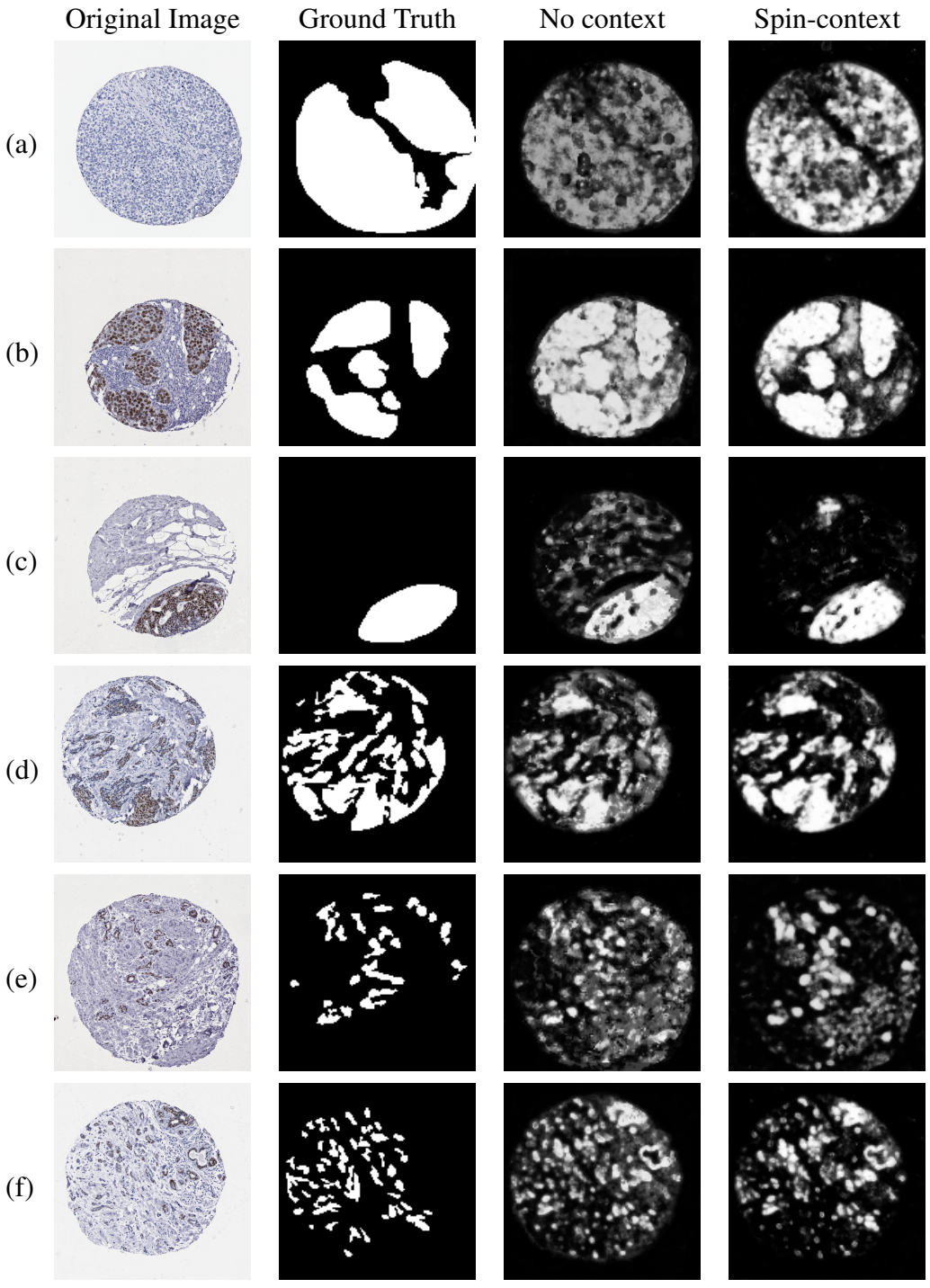


FIGURE 5.5: Subset of results achieved using spin-context. For each TMA spot, images are shown for the ground truth annotation, no context and iteration 6 of spin-context (columns, left to right).

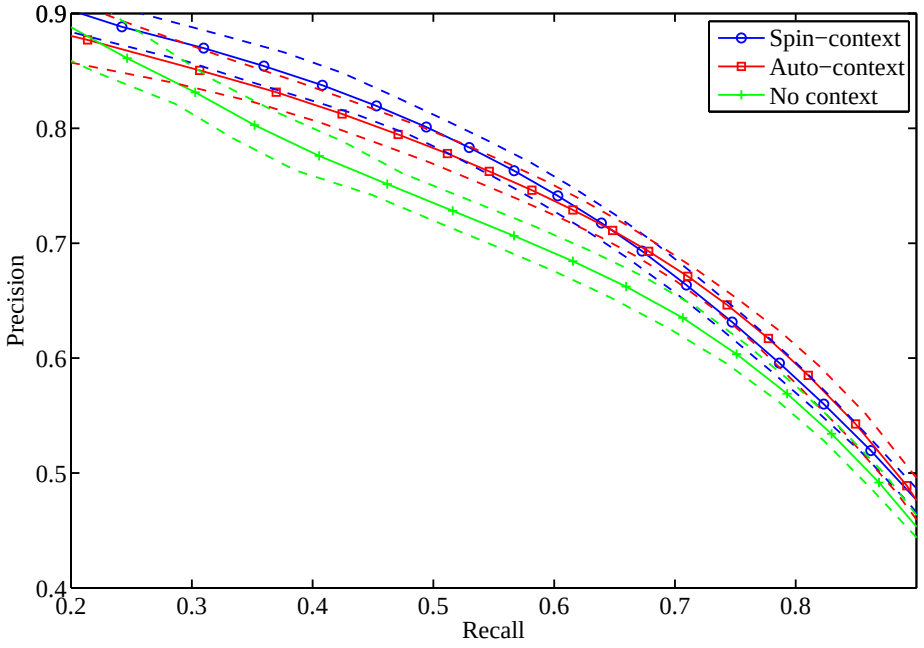


FIGURE 5.6: Precision-recall curves for auto-context and spin-context (3 iterations), and no context (i.e. iteration 1) on MLP classifiers.

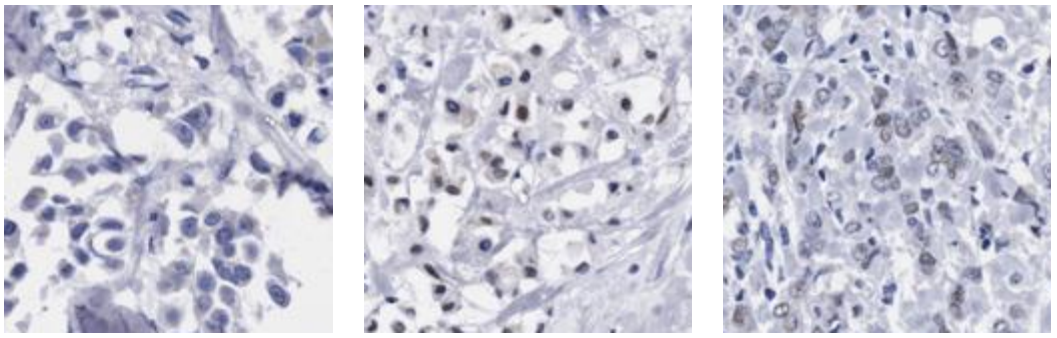
		Pathologist A	
		T	N
Spin-context (A)	T	**0.184**	0.104
	N	0.063	**0.649**

TABLE 5.2: Normalised contingency table comparing spin-context classification maps and pathologist A's segmentation masks.

		Pathologist B	
		T	N
Spin-context (B)	T	**0.187**	0.108
	N	0.068	**0.637**

TABLE 5.3: Normalised contingency table comparing spin-context classification maps and pathologist B's segmentation masks.

when trained on pathologist A and pathologist B, respectively. As in Chapter 4, only grid locations within TMA spot boundaries were evaluated. The inter-rater agreement for spin-context trained on pathologist A and pathologist B was 0.833 and 0.824, respectively. On average, agreements between spin-context and manual segmentation masks were slightly lower than inter-rater agreement between pathologists.



(a) (b) (c)

FIGURE 5.7: Image patches which were incorrectly classified as healthy in spin-context.

On average, 62% of disagreements between spin-context and manual segmentations correspond to false negatives: almost double the number of false positives. In some TMA spots, this can be explained by lumen encased within tumour regions and can arguably be labelled as healthy; in the majority of cases, misclassified tumour regions appear similar to healthy cells. Examples of misclassified tumour image patches are shown in Figure 5.7. In particular, in Figure 5.7(b), abnormal cell development resulted in small cell nuclei scattered within a single tumour region marked by both pathologists. Tumour appearances of this kind was only observed once in the dataset. Therefore it is anticipated more training examples of this kind will improve performance.

5.7.2 Boundary sensitive spin-context

A second experiment which evaluated boundary sensitive spin-context (Section 5.5) is described in this section. F1 measures, the harmonic mean of precision and recall, computed after the removal of context grid locations outwith TMA spot regions are shown in Table 5.4. Standard deviation between repeated experiments were on average 0.007 (iteration 1: $\pm$0.006; iteration 2: $\pm$0.006; iteration 3: $\pm$0.008). In all iterations, boundary sensitive spin-context showed improvement.

To evaluate performance of boundary sensitive spin-context at the spot boundary, masks were created to isolate ROIs (Figure 5.8). The mask shown in Figure 5.8(b)

	Iteration (t)		
	1	2	3
BS spin-context	0.655	0.676	**0.682**
Spin-context	-	0.670	0.676

TABLE 5.4: F1 measures of boundary sensitive (BS) spin-context and spin-context. Iteration 1 does not incorporate spin-context and is therefore identical for both techniques.

(a) (b) (c)

FIGURE 5.8: (a) Spin-context classification map, $\mathbf{p}_n^t$, with 136x136 grid locations, (b) a mask generated to highlight the TMA spot boundary and (c) result after convolution of mask and classification map.

was generated by dilating and eroding TMA spot segmentations by 75 pixels (i.e. 3 grid locations). A subtraction was performed between the dilated and eroded image, resulting in an outline of the TMA spot boundary. Resulting masks were convolved with probability classification maps produced by boundary sensitive spin-context, resulting in tumour probabilities at only the spot boundary (Figure 5.8(c)).

Table 5.5 shows F1 measures computed from ROIs, denoted by the spot boundary mask. The use of context within TMA spot boundaries after one iteration of boundary-sensitive spin-context (iteration 2) is comparable to the result achieved after two iterations of the original implementation of spin-context. This suggests the use of TMA spot boundaries can reduce computational costs associated with spin-context by reducing the number of iterations.

	Iteration (t)		
	1	2	3
BS spin-context	0.644	0.663	**0.671**
Spin-context	-	0.655	0.664

TABLE 5.5: F1 measures of boundary sensitive (BS) spin-context and spin-context in TMA spot boundaries.

5.8 Summary

In this chapter an extension to auto-context called spin-context was described, which captures contextual information from labelled probability maps in a rotation invariant manner. Spin-context was evaluated for the task of tumour localisation on a dataset of ER-stained TMAs. Results showed iterative context extraction adopted in spin-context improved performance, compared to a method which did not incorporate context. Improvement was most noticeable in the first three iterations, after which results converged. Whilst inter-rater agreement between spin-context and pathology experts ($\kappa = 0.829$) were not as high as inter-rater agreement between experts, results are promising. Compared to auto-context, spin-context showed similar outcomes with slightly higher performance at lower recall values.

Spin-context was extended to reduce background interference, by eliminating context locations outwith TMA spot boundaries during training. Results showed that the technique of boundary sensitive spin-context improved performance in all iterations. In particular, boundary sensitive spin-context showed some improvement at the TMA spot boundary, matching performance observed in the previous iteration with the original spin-context implementation.

Chapter 6

RISP: Rotation Invariant Superpixel Pyramid

In the previous chapter, a method which utilised pixel-level image features for tumour segmentation was described. Whilst pixel-level features capture detailed textural information, they fail to capture essential structural information in the tissue. In this chapter a *superpixel* classification method is described which retains information about visual structures such as cellular compartments, connective tissue, lumen and fatty tissue.

Superpixels are described in more detail in Section 6.2 followed by a review of related work. A description of the proposed feature representation called Rotation Invariant Superpixel Pyramid (RISP) is given in Section 6.3. In RISP, a multiscale representation of the tissue is captured which encompasses superpixel geometric, photometric and second-order features (Section 6.3.1). To incorporate information from surrounding superpixels, annuli are adopted to capture frequency and spatial positioning of superpixel visual words (Section 6.3.2, Section 6.3.3).

In Section 6.4, a novel framework called Contextual RISP (CRISP) is described in which image-level and context-level RISPs are combined. Structural and contextual information is captured in an iterative manner, similar to spin-context. Results are

FIGURE 6.1: Superpixel images generated using SLIC code available from http://ivrl.epfl.ch/research/superpixels.

reported in Section 6.7, including comparisons of RISP with related superpixel classification algorithms and manually-obtained tumour segmentation masks.

6.1 Superpixels

A superpixel can be described as a "perceptually meaningful atomic region" of pixels [1]. The characteristics of a superpixel may refer to colour, texture and/or shape depending on the algorithm used to generate superpixels. In the literature, a superpixel image is often referred to as an over-segmentation of an image. Figure 6.1 shows superpixel images generated using SLIC [1] where each superpixel is outlined in black. For each example, different numbers of superpixels have been generated. When comparing superpixels directly, they are compact and roughly the same size, however can freely adapt to nearby boundaries. Within regions which are uniform in colour and/or texture, superpixels adopt a grid-like state.

In a study reported by Neubert and Protzel [105] various properties of state-of-the-art superpixel generation algorithms were compared. The most common technique for generating superpixels is growing from an initial set [1, 86, 139, 141, 144]. Earlier work using the watershed approach [144] performed gradient ascent starting from

local minima to produce watersheds, lines that separate "catchment basins". More recently, in SEEDS [139], superpixels were iteratively produced by performing block-level and pixel-level updates. Block level updates performed a coarse but fast over-segmentation and pixel-level updates refined superpixel boundaries thereby providing a quick but representative superpixel image.

An alternative is to use a graph-based approach. Normalised Cuts by Shi and Malik [128] recursively partition a graph of pixels using contour and textural properties. Felzenszwalb and Huttenlocher [48] proposed a pairwise region comparison algorithm which clustered pixels at each node of a graph; each node was the minimum spanning tree of constituent pixels and was then representative of a superpixel.

In this thesis, Simple Linear Iterative Clustering (SLIC) [1] was used to generate superpixels. SLIC is based on an iterative k-means clustering algorithm which uses the *Lab* colour space and distance between cluster centres to iteratively generate superpixels. Due to its simplicity, SLIC is computationally efficient. However, any of the superpixel generation methods described above could be substituted for SLIC in reported methods.

6.1.1 Motivation

One of the main benefits of using superpixels in image or video analysis is that millions of pixels can be reduced to only a few hundred superpixels thus improving computational complexity. For analysis of high resolution images, which is often the case in digital pathology, this is a desirable property.

However superpixels also provide additional structural properties which are important when modelling tissue. In histopathology, tissue structure is complex, and as shown in the previous chapter, is difficult to model from low-level pixel features. Superpixels, on the other hand, can adapt to surrounding tissue thereby providing rich, descriptive features. Figure 6.2 shows superpixels generated in breast histology images. Note that the tissue structure is indirectly captured without explicit semantic segmentation. For

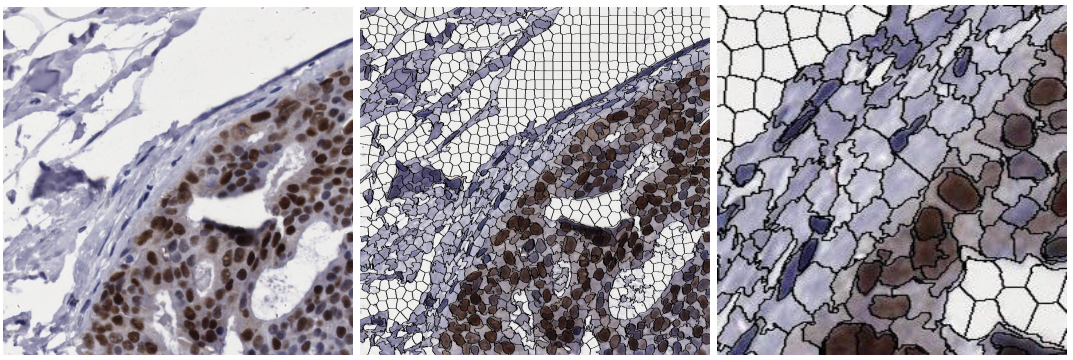


FIGURE 6.2: Image patch of breast tissue stained with Haematoxylin and ER (left), its corresponding SLIC superpixel image (middle), and a magnification of superpixels, some of which encase cells (left).

example, superpixels elongate to fit to lymphocytes and fibroblasts in stromal regions but remain compact in regions containing lumen.

To select a suitable value for the number of superpixels per image, Z, an experiment was designed to explore the properties of the superpixel image when this parameter was varied. In Figure 6.3 each superpixel is replaced by the average RGB colour value of all pixels contained within that superpixel. When Z is low, intermediate tissue structure (e.g. lobules) are retained however smaller cellular components are lost, particularly clustered epithelial cells and fibroblasts in connective stromal tissue. Larger structures such as fat are still retained, however the boundaries are somewhat less refined. When Z is high, the superpixel image closely resembles the original image and very little information is lost between the pixel-level and superpixel-level representations. An expert pathologist considered $Z = 50,000$ to retain tissue structure so that tumour regions were clearly distinguishable. At this setting, two or more superpixels were often used to represent epithelial nuclei.

6.2 Related work

Recently, the usage of superpixels in the computer vision literature has grown so that they are not only used to reduce computational costs, but also provide descriptive information in addition to or instead of pixel-level features.

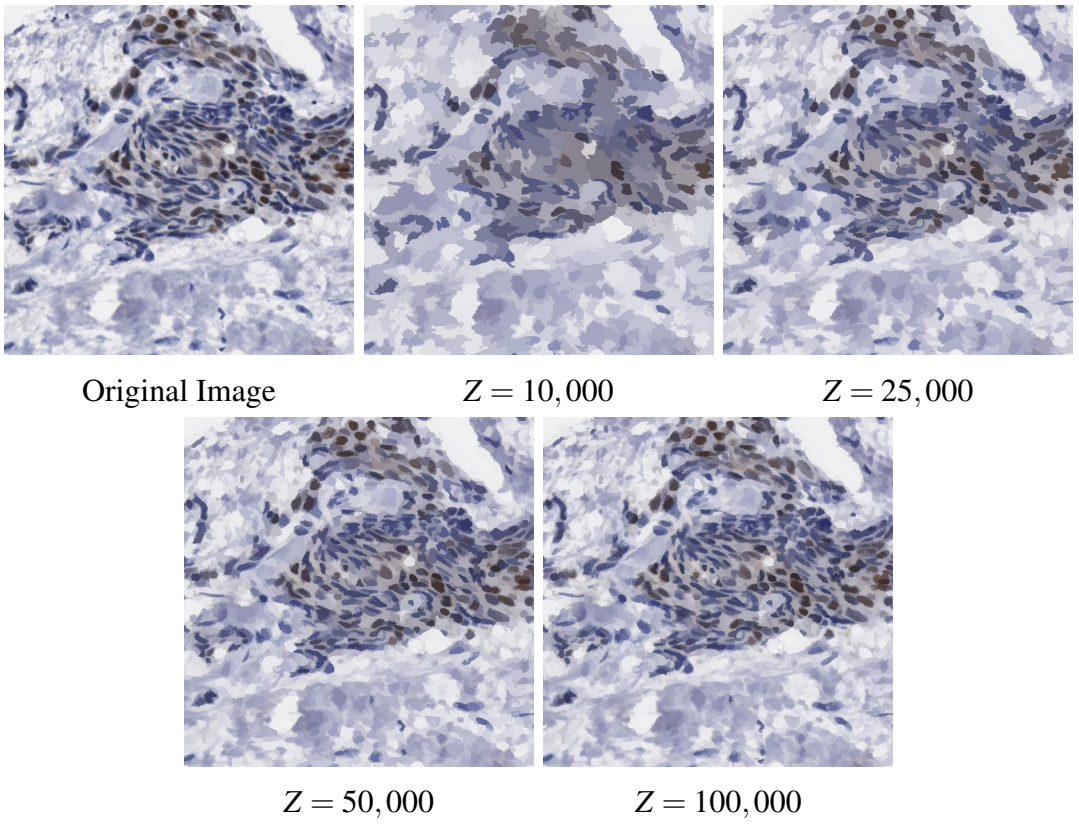


FIGURE 6.3: SLIC superpixel representation when number of superpixels, Z, is varied. Each superpixel is rendered with the average RGB value of pixels it contains.

A popular use of superpixels is to map them onto a Conditional Random Field (CRF) model thereby retaining the structure of the superpixel image in relation to neighbouring superpixels. Fulkerson *et al.* [51] showed the benefits of using a CRF model with a SIFT bag-of-words algorithm for object localisation. Similarly, Li and Sahbi [89] constructed CRFs by analysing superpixels generated at multiple resolutions in which the pairwise term took into account neighbouring superpixels in four directions. Hao *et al.* [60] segmented tumours in ultrasound images using a combination of nested classifiers and CRFs. In the above examples, CRFs were used as a refinement process to obtain precise boundaries after extraction of superpixel appearance, shape and/or morphometric features. CRFs are not used in this thesis but it would be interesting to explore their use in future work to capture context from learned classification maps (Section 9.2).

In recent work, superpixels have been used to capture context in various innovative forms. Gould *et al.* [57] proposed a relative location prior based on an underlying superpixel representation, which provided a general class probability map thereby encoding surrounding objects. Dickinson *et al.* [37] adopted a perceptual growing approach by analysing superpixel graphs at multiple scales, wherein, at each scale, graphs were compared for symmetrical parts. Similarly, Kumar and Hebert [79] used a multi-scale approach to model interactions between superpixels and "sub-superpixels", where sub-superpixels are superpixels generated from a single superpixel in the preceding layer. Techniques for grouping or splitting superpixels at various scales have also been employed in other methods [77, 86]. Previous work has shown that context in the form of relationships between superpixels, whether at a single scale or multiple scales, is important for capturing descriptive feature representations. Later in this chapter, context from surrounding superpixels is explored further with the aim of capturing complex tumour patterns.

In histopathology, Gorelick *et al.* [56] extracted textural and appearance features from superpixel images of prostate cancer and trained them using Adaboost. Pixel-level features were then incorporated in a rotation invariant manner to provide contextual information. A similar technique is adopted in this chapter. However, instead of pixel-level features, superpixels are adopted to capture context with the aim of providing a richer context descriptor. In other work, Beck *et al.* [18] built an explicit stroma versus epithelial superpixel classifier to identify cell nuclei in breast tissue. In [18], several relational and morphological features were extracted from cell nuclei and stromal regions to predict survival. It was shown that features extracted from stromal cells were better predictors of patient survival than features extracted from epithelial cells, contradictory to the current approach for manual analysis of histopathology images.

6.3 Method

In the following sections, a series of representations for modelling superpixels in a rotation invariant manner is described: Bag-of-Superpixels (BoS), Spatial Bag-of-Superpixels (S-BoS) and RISP. Before doing so, superpixel features used to construct visual words are defined.

6.3.1 Definition of superpixel features

As mentioned previously, SLIC [1] was used to construct superpixels. Here, superpixels are denoted as $\mathbf{s} = [s_1, ..., s_Z]$ where Z is the total number of superpixels. The number of superpixels was assigned so that the area of a single superpixel rarely exceeded the area of a cell nucleus. The average area of the extracted superpixels was 221 pixels with a standard deviation of 34 pixels. Whilst subcellular superpixels may not capture rich textural properties, geometric properties extracted from these superpixels indirectly model cell shape, size etc. without explicit cell segmentation.

For each superpixel, a set of features was obtained to describe its appearance and geometric properties. To capture additional textural information within each superpixel, Haralick [61] features were adopted. Compared to differential invariants utilised in Chapter 5, Haralick features can be computed quickly for thousands of superpixels. Second-order features, including the number of superpixel neighbours and variance between neighbouring arc lengths, were also included in the feature representation. Here, the variance between neighbouring arc lengths is a measure of the proportion of the superpixel perimeter shared between neighbouring superpixels. For example, an elongated neighbouring superpixel is more likely to share a larger proportion of the shared perimeter compared to a more compact neighbouring superpixel. Superpixel features are outlined in Table 6.1. The resulting superpixel features were normalised and concatenated to form a descriptor, $\mathbf{f}_z$, per superpixel.

Appearance	Mean red, green and blue values of containing pixels Mean and variance of greyscale values of containing pixels 13 Haralick texture features ($f_1 - f_{13}$); angular second moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measures of correlation
Geometry	Compactness of superpixel contour Eccentricity of superpixel contour Area of superpixel i.e. number of containing pixels Perimeter i.e arc length of superpixel contour
Neighbours	Number of immediate superpixel neighbours Variance between neighbouring arc lengths

TABLE 6.1: List of features extracted from each superpixel.

6.3.2 Bags-of-Superpixels

To perform superpixel classification, one approach is to use the superpixel features outlined in Table 6.1 directly. However this produces poor results due to lack of context. As discussed in Section 6.2, capturing neighbouring superpixels is important for essential contextual information. As such, extracted superpixel features were used in the bag-of-words framework. Superpixel descriptors were quantized using a K-means dictionary, $\mathbf{C}$, to produce a set of visual words. A circular window with radius R was then positioned at the centre point of the superpixel to be classified. Visual words of superpixels within the circular window were histogrammed, resulting in a Bag-of-Superpixels (BoS). The BoS representation described here shares similarities to the colour and spatial superpixel-based BoW representation proposed by Shu *et al.* [129].

6.3.3 Spatial Bags-of-Superpixels

Whilst BoS is a simple yet powerful representation, its main drawback is lack of spatial information. In breast cancer, spatial information of cell nuclei is an important property for modelling breast cancer. For example in later stages of DCIS, cancer cells are tightly packed within a duct whereas in a healthy duct epithelial cells are arranged in a ring-like pattern.

Algorithm 2 Spatial Bag-of-Superpixels (S-BoS)

Input: Image $\mathbf{x}_n$, radius R (pixels), number of annuli Q, superpixel codebook $\mathbf{C}$, number of superpixels Z

Run SLIC on $\mathbf{x}_n$ to generate superpixels, $\mathbf{s} = [s_1 ... s_Z]$

Extract superpixel features, $\mathbf{F} = \{\mathbf{f}_1 ... \mathbf{f}_Z\}$

for *each superpixel,* s_z, *in* **s do**

  Identify superpixels indexed by **t** within circular window with radius, R, centred at $c(s_z)$

  Initialise S-BoS histogram, H_z

  for *each superpixel,* t_y, *in* **t do**

    Lookup codeword, v_y for $\mathbf{f}_{t_y}$ in **C**

    Compute $d = ||c(t_y) - c(s_z)||$, c returns the centre point of a superpixel

    Increment $H_z(v_y, \left\lfloor \frac{Qd}{R} \right\rfloor)$

  end

  Normalize H_z

end

To capture this information, the spatial distribution of each visual word is modelled in a spatial Bag-of-Superpixels (S-BoS) histogram. Spatial information is captured in the form of equally spaced annuli within the circular window utilised in BoS. The method for constructing S-BoS histograms is outlined in Algorithm 2. $H_z, z \in 1...Z$, denotes a S-BoS histogram incorporating Q annuli. d is the distance between two superpixel centre points returned from function $c(\cdot)$.

This approach has some similarities to [56]. However, in this work, neighbouring superpixels are analysed from their visual words instead of pixel-level RGB values. In doing so, important structural information from the underlying superpixel representation is retained.

6.3.4 Rotation Invariant Superpixel Pyramid

Spatial pyramids, originally proposed by Lazebnik *et al.* [83], partition an image repeatedly to compute a bag-of-words histogram per cell or sub-region. In [83], sub-regions consisted of square grids and, appropriately for the intended application, the

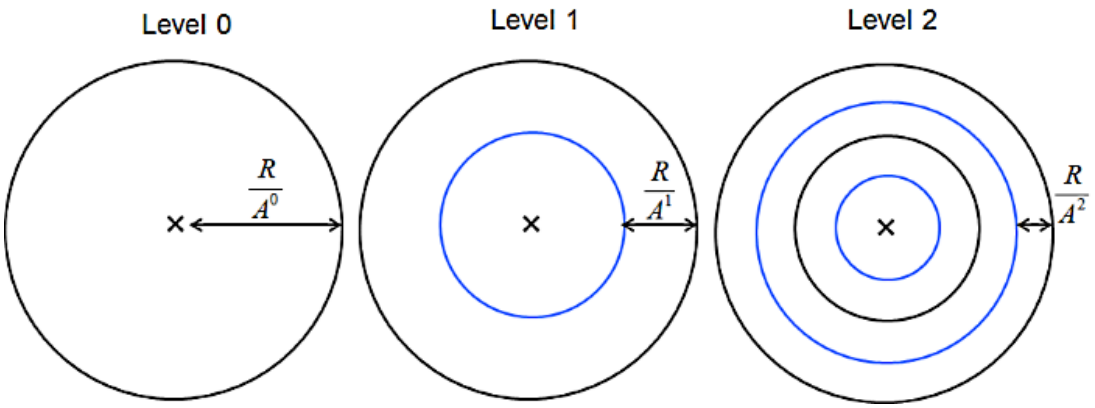


FIGURE 6.4: Levels 0, 1 and 2 of RISP. The BoS representation is level 0 after which partitions are applied iteratively according to A.

representation was not rotation invariant. Here, a novel Rotation Invariant Superpixel Pyramid (RISP) is proposed in which S-BoS histograms are computed per level (Figure 6.4). The number of annuli grows exponentially with A, a growth factor, in each level. In reported experiments, $A = 2$; therefore each annulus in level L, is replaced by two annuli in $L+1$. By utilising circular annuli, histograms computed from each annulus are rotation invariant. The final RISP context descriptor is the concatenation of S-BoS histograms for each level to form a multiscale representation. To provide complementary local features, the superpixel descriptor for the centre superpixel is concatenated with each RISP.

Previous work by Wiliem *et al.* [150] used a technique similar to this for cell classification. However, here a general-purpose approach is adopted, applicable to other computer vision tasks. In RISP, spatial information is extracted implicitly without the need for explicit cell segmentation as in [150].

6.4 Contextual RISP

Whilst RISPs can be classified directly, in this section an alternative framework called Contextual RISP (CRISP) is proposed. CRISP is an adaptation of spin-context described in Chapter 5, whereby posterior probabilities from superpixel classification maps are captured in an iterative manner. In CRISP, image features take the form

of *image-level* RISPs. Image-level RISPs are equivalent to the RISP representation described earlier (Section 6.3) concatenated with superpixel features for the central superpixel. To capture contextual information, *context-level* RISPs are proposed. Context-level RISPs reflect posterior probabilities within superpixel classification maps. Before describing CRISP, context-level RISPs are defined.

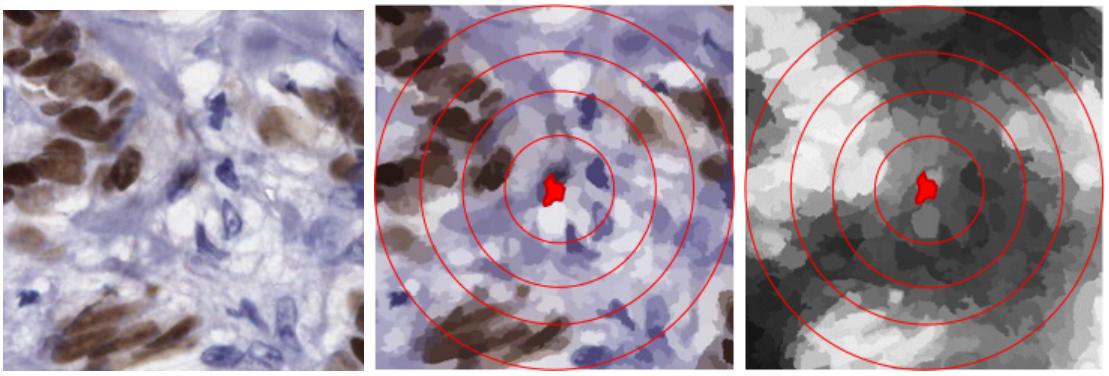
6.4.1 Context-level RISP

A context-level RISP takes a similar form to the image-level RISP. However instead of superpixel visual words, frequencies of superpixels' posterior probabilities generated in iteration $t-1$ in the form of a learned superpixel classification map, are encoded. An illustrative comparison of image-level and context-level RISPs is shown in Figure 6.5. Whilst the RISP structure is identical, the baseline representation from which RISPs are constructed is altered. In the context-level RISP, tumour probabilities assigned to superpixels during testing are modelled in a rotation invariant manner.

As in RISP, a circular support with radius R_c is centred on the superpixel to be classified. Posterior tumour probabilities from classified superpixels within the circular window contribute towards the context-level RISP representation. In each pyramid level, S-BoS histograms are constructed where each row denotes an annulus and columns represent probability distributions. Frequencies of posterior probabilities are captured in B equally-spaced bins. S-BoS histograms from each pyramid level are concatenated to form the final context-level RISP representation.

To retain information about the central superpixel, the tumour posterior probability assigned to the central superpixel in iteration $t-1$ is also appended to each context-level RISP.

An overview of the CRISP framework is shown in Figure 6.6, where $\mathbf{p}_n^{(t)}$ is the superpixel classification map produced in iteration t, for image $\mathbf{x}_n$.



Original Image Image-level RISP Context-level RISP

FIGURE 6.5: Illustrative comparison of image-level and context-level RISP. Image-level RISPs are constructed from superpixel visual words, illustrated by their average RGB values, whereas context-level RISPs are constructed from posterior probabilities.

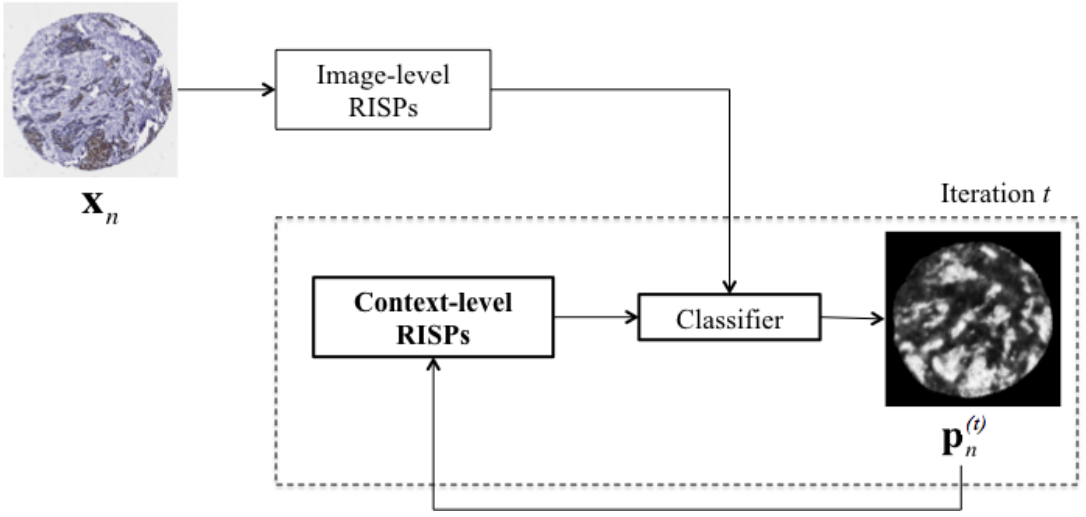


FIGURE 6.6: Overview of Contextual RISP.

For each iteration t, image-level RISPs are concatenated with context-level RISPs constructed from the classification map generated in the previous iteration, $\mathbf{p}_n^{(t-1)}$. The output for iteration t is an updated superpixel classification map, $\mathbf{p}_n^{(t)}$. As a result, a series of T superpixel classification maps are produced per image. As there is no context available in iteration 1, context descriptors consist of a uniform prior.

By utilising both image-level and context-level RISPs, structural information from the original image is retained and complements contextual information extracted from classified superpixels. Compared to spin-context, context-level RISPs give a richer

representation of context within the context window at multiple scales. Additionally, the context-level RISP is a general-purpose model that can be tailored to various computer vision tasks with ease.

6.5 Nested cross-validation

During training, it is essential that training and test data are kept separate throughout the CRISP framework (i.e. across multiple context iterations). To ensure separation of training and test data, a stacked generalisation approach described by Jampani *et al.* [68] was adopted. In [68], the training set was partitioned to construct auto-context descriptors across multiple context iterations without compromising a held-out test set. This work is an extension of the stacked generalisation technique [151]. The traditional cross-validation framework is adapted to partition training folds, as described in [68]. This is referred to as a *nested* cross-validation setup. Here, a training set refers to a subset of the dataset used to train a classifier. A validation set refers to a subset of the dataset used to validate a trained model.

Let S denote a set of labelled data. In a traditional U-fold cross-validation setup, $S = \bigcup_{u=1}^{U} S_u$, where sampling data associated with each fold is indexed by u. In fold u, S_u is reserved for validation whilst the remaining set, shown in (6.1), is used to train a model.

$$\bar{S_u} = \begin{cases} \bigcup_{i=1}^{U-1} S_i, & \text{if } u = U \\ (\bigcup_{i=1}^{u-1} S_i) \bigcup (\bigcup_{j=u+1}^{U} S_j), & \text{otherwise} \end{cases} \tag{6.1}$$

In the nested cross-validation framework, $\bar{S_u}$ is further split into V sub-folds, where sampling data associated with each sub-fold is denoted by G_v,

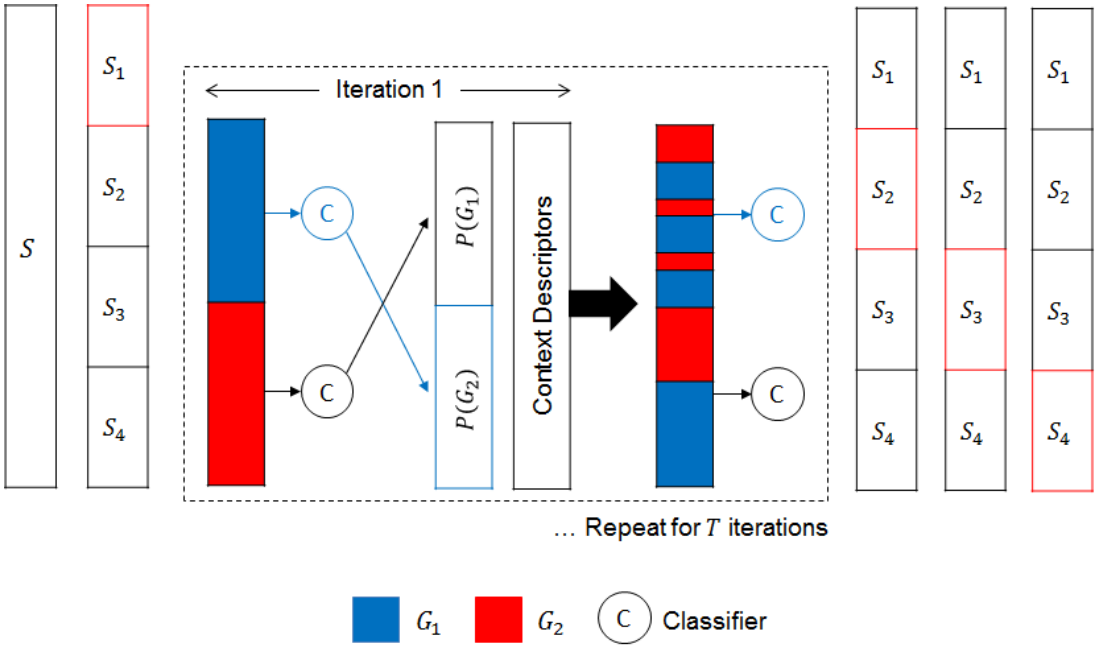


FIGURE 6.7: Split of train and validation data in a nested cross-validation setup ($U = 4$, $V = 2$). Predictions, P, are obtained from trained models produced in each sub-fold.

$$\bar{S}_u = \bigcup_{v=1}^{V} G_v \tag{6.2}$$

The validation set in each sub-fold is G_v. Remaining data in $\bar{S}_u$ is reserved for training. Figure 6.7 shows a nested cross-validation setup where $U = 4$ and $V = 2$. Sub-folds are assigned randomly and are shown in red (G_1) and blue (G_2). Predictions generated in each sub-fold are used to construct context descriptors for the following context iteration.

To produce predictions on the test fold, i.e. S_u, one model is trained using set $\bar{S}_u$. In iteration t, context descriptors generated in sub-folds (together with image-level RISPs) form the input to the classifier. This produces a trained model from set $\bar{S}_u$. Posterior probabilities for set S_u are then retrieved from the trained model. S_u has no influence during construction of the trained model and is therefore representative of a hidden test set. The above process is repeated for U folds.

6.6 Experiments

Tumour localisation was evaluated using 8-fold cross-validation on a dataset of 32 ER-stained TMAs (Chapter 4). In CRISP, nested cross-validation experiments are reported with $U = 8$ and $V = 2$. In reported experiments, labels retrieved from pathologist A were used for evaluation purposes. A balanced training set was used with identical numbers of positive and negative samples, randomly sampled. A linear SVM classifier was implemented in the LIBLINEAR [45] framework. Posterior probabilities were obtained using Platt's method [117] which maps SVM outputs between 0 and 1 by fitting to a sigmoid function with two trade-off parameters. Parameters are calibrated using maximum likelihood estimation. To avoid overfitting, Platt's method uses a 5-fold cross-validation setup. A grid search was performed to obtain an optimal cost parameter. $50,000$ SLIC superpixels were extracted from each TMA spot image. For a 3600 x 3600 pixel image, $50,000$ SLIC superpixels can be computed in 29 seconds on an Intel i5-2410M 2.3GHz processor.

Image-level RISPs were constructed using a circular window with radius, $R = 100$ pixels. For context-level RISPs, R_c was varied to determine the impact of the size of the context window. RISP level parameters ($L = 3$, $A = 2$) were identical for image-level and context-level RISPs.

TMA spots were segmented manually to reduce background interference. Superpixels with centre points within spot boundaries were classified in reported experiments.

6.7 Results

6.7.1 RISP

Figure 6.8 shows ROC curves for the following implementations,

- The proposed **RISP** representation.

- Iteration 3 of **spin-context** as described in Chapter 5 using SVM classifiers.
- Spatial Bags-of-Superpixels, **S-BoS**, for 2 and 4 annuli.
- Bags-of-Superpixels, **BoS**.
- An implementation of a superpixel classification method described by **Gorelick** *et al.* [56] which uses pixel-level features to incorporate context information.
- Superpixel **autocorrelograms** as described in Appendix A which model spatial distributions between pairs of superpixels.
- Superpixel **features** with no context.

There was a small improvement in classification performance between BoS and S-BoS with 2 annuli, showing the benefits of incorporating spatial information. 3-level RISP which incorporated 7 $(1+2+4)$ annuli also showed a noticeable improvement compared to S-BoS; however increasing the number of annuli in S-BoS resulted in minor improvement. Compared to Gorelick's implementation [56] and superpixel autocorrelograms, RISP was superior. Superpixel features alone with no contextual information showed the worst performance, enforcing the importance of incorporating contextual information from surrounding superpixels.

ROC curves shown in Figure 6.8 might suggest spin-context is superior to RISP (although it does not dominate). In order to compare classification performances from ROC representations, an alternative analysis metric called the cost curve [43] is used which enables classification performance (i.e. normalised expected cost) to be observed with respect to class distributions. Each slope in the cost space directly corresponds to a single point in the ROC space, whereby a slope is the line from $(0, FP)$ to $(1, 1-TP)$. The resulting cost curve is the lower envelope of those lines in the cost space. Equal costs were adopted for positive and negative training samples therefore the y-axis in reported curves denotes error rate. The x-axis shows a full range of tumour distributions where $x = 1$ means all examples are tumour and $x = 0$ means all examples are non-tumour.

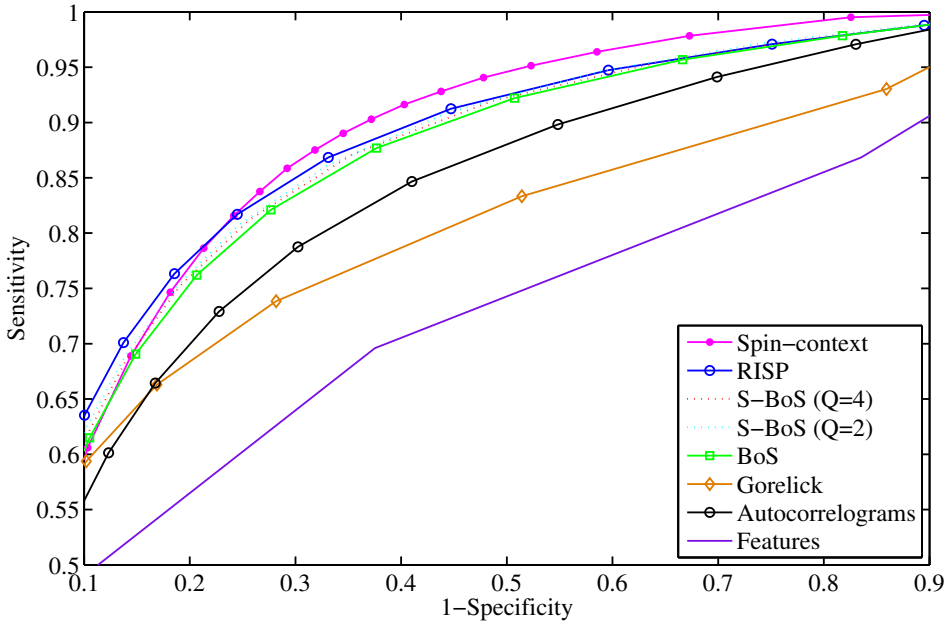


FIGURE 6.8: ROC curves for spin-context, RISP, BoS, S-BoS, superpixel features (Features), method as described in [56] (Gorelick) and superpixel autocorrelograms (autocorrelograms) with 200 codewords.

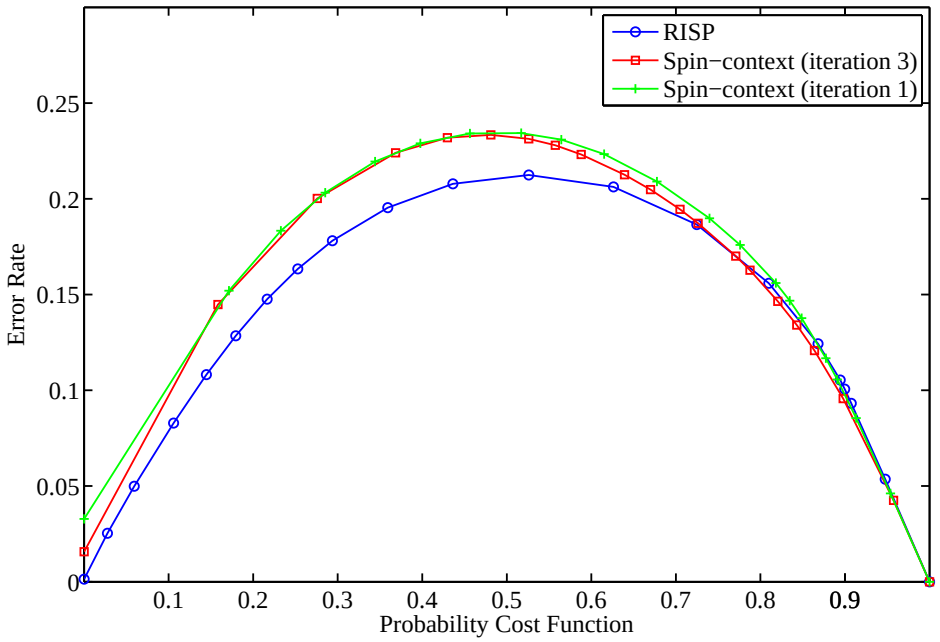


FIGURE 6.9: Cost curves for RISP, and iterations 1 and 3 of spin-context.

		Number of Codewords			
		25	50	100	200
BoS		0.623	0.662	0.680	0.688
RISP	2 levels $\{0,1\}$	**0.650**	0.679	0.671	0.700
	3 levels $\{0,1,2\}$	0.648	**0.681**	**0.692**	**0.703**

TABLE 6.2: F1 measures for BoS, S-BoS and RISP.

The cost curves for RISP and spin-context are shown in Figure 6.9. From reported cost curves, RISP showed lower error rates than spin-context across lower tumour distributions, with similar performance at higher tumour distributions. Specifically, when considering the proportion of positive samples in the evaluation dataset (31%), RISP yields an error rate of $\sim$17.8%. At this point on the x-axis of the cost curve, RISP surpassed three iterations of spin-context and showed considerable improvement to pixel-level features – spin intensity features [82] and differential invariants [126] – adopted in Chapter 5.

Table 6.2 shows F1 measures for RISP for various numbers of codewords. As the number of codewords increased and as more levels were incorporated in RISP, the accuracy continued to increase; differences between 100 and 200 codewords were marginal. At lower dictionary sizes, RISP was more effective at lower levels. Compared to BoS (i.e. RISP level 0), RISP showed improvement for the majority of reported dictionary sizes.

Table 6.3 shows AUC and F1 measures when the compactness of generated superpixels was varied. A low compactness value (i.e. very compact superpixels) gave the worst performance as superpixels were less distinguishable between regions containing different tissue types. However at the other extreme, generated superpixels had noisy boundaries; the complex textural appearance in the pixel-level image resulted in highly variable superpixels of different shapes and sizes. A balance between these two properties resulted in best performance.

	Compactness			
	1	2	5	10
AUC	0.836	0.847	**0.859**	0.851
F1	0.675	0.684	**0.703**	0.694

TABLE 6.3: AUC and F1 measures achieved when superpixel compactness was varied in SLIC.

6.7.2 CRISP

A second experiment was designed to evaluate CRISP. Image-level RISP parameters were fixed to determine the impact of varying context in terms of the context support window size.

A comparison between CRISP, RISP and spin-context is shown in Figure 6.10 in the form of cost curves [43]. Here, the size of the context-level RISP window was 100 pixels. CRISP showed a strong performance gain compared to spin-context across all tumour distributions. Compared to RISP, error rates were similar performing slightly better at higher tumour distributions (0.50-0.75) and slightly worse at lower tumour distributions (0.25-0.45).

A visual representation of superpixel classification maps produced in CRISP is shown in Figure 6.12. Results varied between TMA spots. In Figure 6.12(a) tumour probability values dropped within annotated tumour regions after iteration 2, suggesting further context was detrimental to classification performance. However in Figure 6.12(b) tumour classification improved with additional context information. Figure 6.12(c) shows one of the worst cases in the reported dataset. Note that initial classification was poor in this case as cancer cells were negatively stained and appeared similar to healthy epithelial cells. Figure 6.12(d) - Figure 6.12(f) show cases in which CRISP slightly improved tumour localisation. In these cases initial classification was strong. As such it may be that RISP already incorporated sufficient context and therefore any additional context had minor impact on overall performance. Future work will explore CRISP further (Section 9.1).

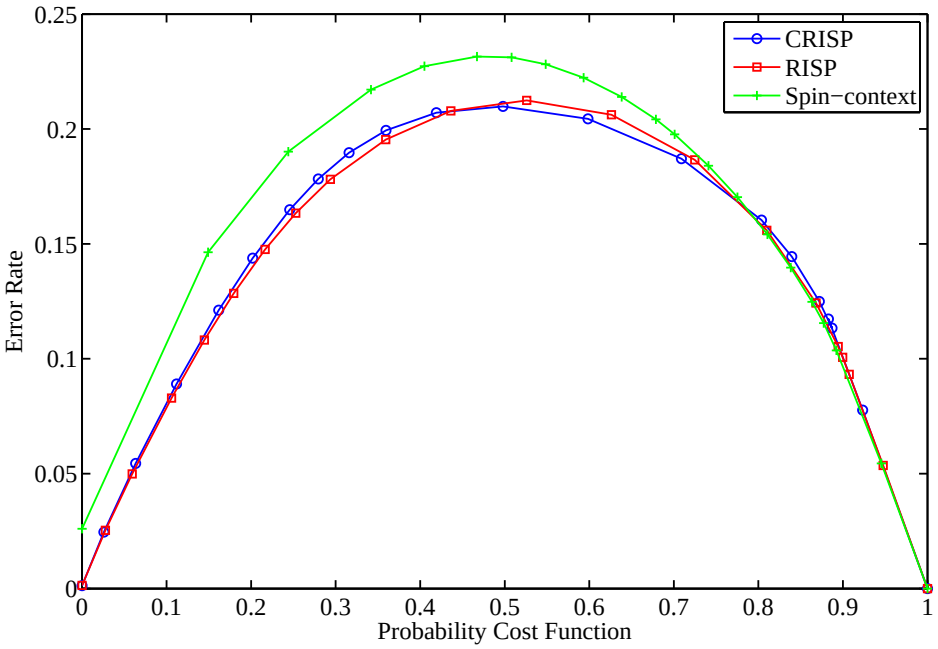


FIGURE 6.10: Cost curves for two iterations of CRISP and spin-context, and RISP.

Figure 6.12 shows ROC curves when the radius of the context window, R_c, was varied. Error rates were similar regardless of the window size suggesting the size of the context window does not greatly impact performance. The cost curves shown in Figure 6.13 demonstrate a larger window size would slightly improve classification accuracy when tumour distributions are high but have the opposite effect if tumour distributions are low.

6.8 Summary

In this chapter, a novel rotation invariant superpixel representation, RISP, was proposed which uses a compact superpixel representation to capture structural information in histopathology images. To incorporate spatial information at multiple scales, a pyramid representation was adopted. At each pyramid level, spatial configuration of superpixel visual words was modelled. Results showed RISP performed favourably compared to related superpixel classification algorithms.

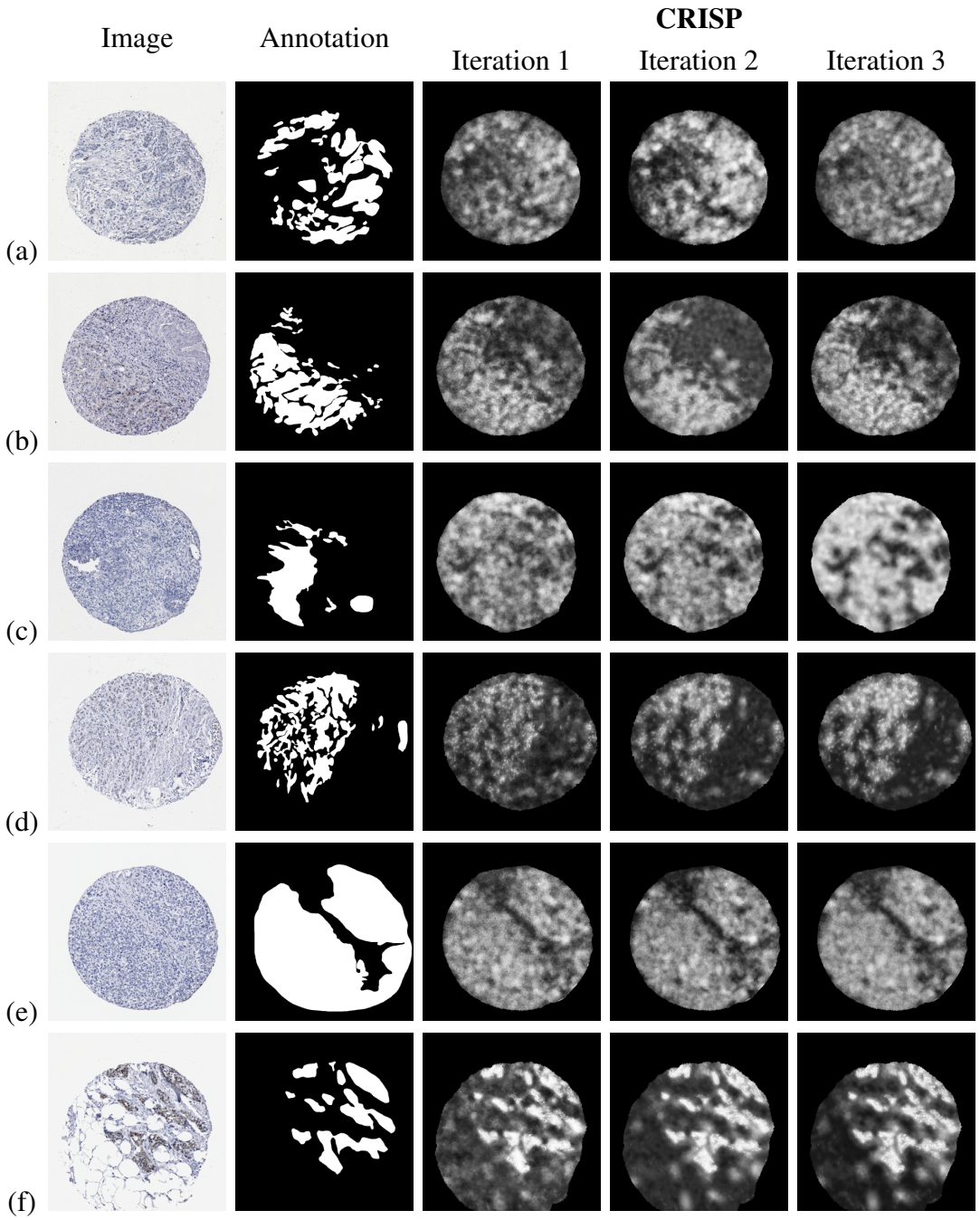


FIGURE 6.11: Histopathology images, manually annotated tumour regions and superpixel classification outputs for iterations 1, 2 and 3 of CRISP (ordered left to right).

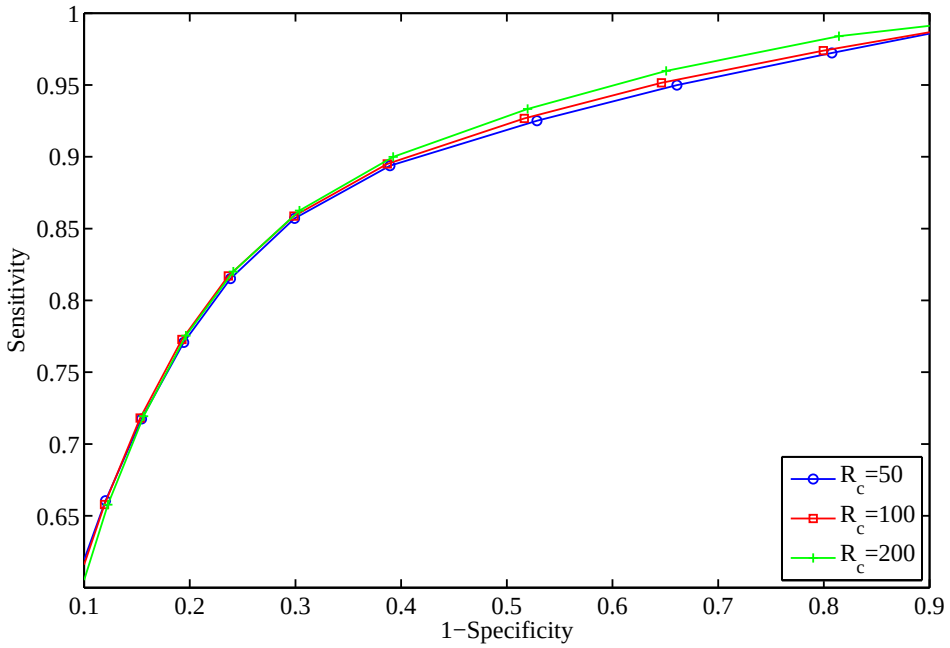


FIGURE 6.12: ROC curves for different sizes of context windows after two iterations of CRISP.

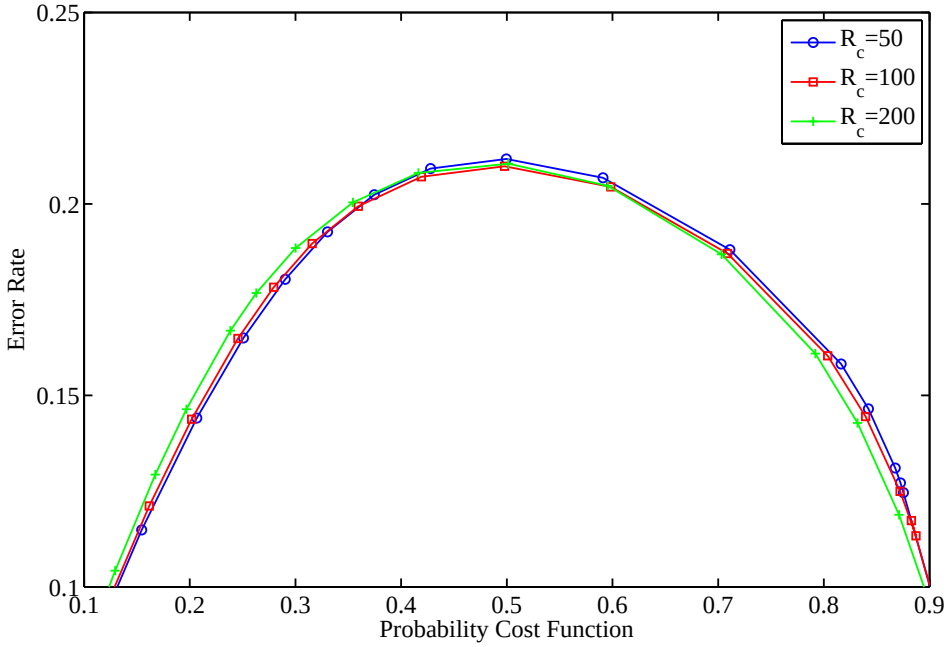


FIGURE 6.13: Cost curves for different sizes of context windows after two iterations of CRISP.

Cost curves, proposed by Drummond and Holte [43], were used as an alternative evaluation metric to the ROC curve. RISP and spin-context cost curves (Figure 6.9) showed error rates were lower using RISP, particularly when tumour distributions were low, as in the current dataset. RISP also showed comparable performance to boundary sensitive spin-context proposed in Chapter 5.

An iterative context framework called Contextual RISP (CRISP) was proposed, in which image-level and context-level RISPs were fused together. Superpixel classification maps produced in each iteration were used to model contextual information in subsequent iterations. A nested cross-validation experiment (Section 6.5) was adopted which ensured separation of training and validation data. Results showed context-level RISPs did not improve results in the reported dataset. Visual interpretation of superpixel classification maps suggests image-level RISPs already encompassed sufficient contextual information and therefore any additional context had minor impact on resulting performance.

Chapter 7

Automated Tumour Localisation: Clinical Impact

In Chapter 4, inter-rater agreements were reported between expert pathologists to gain some insight into the current "gold standard" for manual tumour localisation. However, accepting that expert specialist breast pathologist review is the gold standard for interpretation of biomarkers such as ER, any automated approach needs to be comparable to and consistent with such expert assessment.

In previous chapters, automated solutions to tumour localisation were explored using ER staining of breast TMAs as a clinically relevant exemplar. In this chapter, manual and automated segmentation masks are compared in an attempt to measure the current performance of automation for clinical assessment. In the reported study, manual segmentations replicated the current manner with which pathologists interacted with a widely used FDA-approved IHC scoring algorithm. Manual input of this form was compared with automated annotations generated by RISP. Specifically, IHC Allred scores [5] and Quickscores [36] extracted from segmentation masks were compared. Furthermore, to measure the impact of IHC in treatment decision-making, cut-offs were applied to extracted IHC scores to label TMAs are either ER+ve or ER-ve.

7.1 Introduction

In previous studies comparing automated and manual IHC scores [46, 123], evaluations were performed based on measurements retrieved after cell analysis was performed at the pixel-level. However the challenge in current digital pathology applications lies in distinguishing healthy from cancerous tissue [58, 116]. In a study reported by Rizzardi *et al.* [123], a pattern recognition system was adopted to localise tumours; however no analysis was reported of the accuracy of automatically-generated tumour boundaries. Cass *et al.* [25] adopted a semi-automated scoring system and concluded accuracy strongly depended on *a priori* identification from experts for training. Despite this, clinical studies show automatically obtained IHC scores concord closely with manual assessment. In a study of $3,484$ TMAs reported by Turbin *et al.* [138], agreements between automated IHC scores aligned closely with inter-rater agreement between experts. Note no tumour localisation was performed in this study; full TMA spots were automatically analysed.

Benefits of automation include standardisation of IHC measurements, which enable IHC scoring that is objective and reproducible. In HER2 assessment of breast TMAs, Gustavson *et al.* [59] demonstrated an automated system yielded 94.8% concordance, standardised across laboratories and operators, concurrent with recommendations submitted by The American Society of Clinical Oncology and the College of American Pathologists. Furthermore, with improvements in processing power and digital storage over the years, machines can analyse digital slides in less time than it would take to manually analyse them, thus showing potential to refocus pathologist expertise to more difficult cases where disease is difficult to identify [92]. Currently statistics show prostate pathologists are spending 80% of their time sieving through benign tissue [58]. This time can be better spent analysing malignant or suspicious cases.

The main advantage of automation is in quantitative analysis of digital slides as measurements can be acquired with greater precision compared to manual analysis. For example, in the case of IHC scoring, manual analysis is performed on an ordinal scale whereby proportion of cells are estimated into five or six categories (Section 2.5.1).

As a machine can scan and analyse digital slides quickly, cell counting is a trivial task which can be performed with greater precision, in most applications down to a single cell. It is important to note, that accuracy of cell counting software relies upon the image analysis algorithm used. Whilst some systems may produce precise measurements, this does not necessarily indicate *accurate* measurements.

In this chapter, a study is reported which measured agreements between manual and automated segmentations of tumour regions. Unlike in previous studies, various stages including annotation of tumour regions, extraction of cell measurements (percentage of positive cells, intensity scores) and computation of IHC scores is evaluated. In doing so, a deeper understanding of the impact of pixel-level disagreements in annotated tumour regions is provided.

7.2 Methods

7.2.1 Automated spot segmentations

Automated segmentation masks consisted of superpixel classification maps produced by RISP, described in Chapter 6. 3-level RISPs were adopted with a growth factor, A, of 2 and a circular support window with a radius of 100 pixels. 200 codewords were used to encode superpixel features. Linear SVM classifiers were adopted, with balanced numbers of positive and negative training examples.

Each of the 32 TMA spots in the dataset were automatically segmented twice: once trained on pathologist A's segmentation mask and once trained on pathologist B's segmentation mask.

7.2.2 ER scoring of segmented spots

The FDA-approved Aperio IHC Nuclear Version 10 algorithm (Aperio Technologies, CA, USA) was used to estimate ER scores based on the automated and manual segmentation masks obtained. Only pixels labelled as tumour were passed to the scoring algorithm. Masks were created by convolving the original image with binary segmentations. Automated binary segmentations were created by thresholding superpixel classification maps at 0.5.

The Aperio IHC algorithm identifies nuclei automatically and outputs a staining intensity score (ranging from 0 to 3) and an estimate of the percentage of positively stained cells. From these measurements, IHC scores (Allred score and Quickscore) were computed for manually and automatically obtained segmentation masks. Comparisons are reported to assess the extent to which differences in these segmentations affected scoring.

7.3 Results

7.3.1 Segmentation comparison

In pixel-level comparison of manually hand-drawn segmentation masks, pathologists differed in their labelling of 9% of pixels (Section 4.3). Comparisons of each pathologist's manual segmentations with those produced automatically are shown in Table 7.1 and Table 7.2, respectively. On average agreements between automated and manual segmentation masks was $\kappa = 0.811$. Despite this being lower than inter-rater κ agreements, results are promising. With further training examples and advancements in image analysis, automation shows potential to replace manual input in clinical trials in the near future.

Distributions of disagreement types (Type 1, Type 2, Type 3) are shown in Table 7.3. A visual interpretation of agreements and disagreement types is shown in Figure 7.1.

		Pathologist A	
		T	N
RISP (A)	T	**0.221**	0.097
	N	0.092	**0.591**

TABLE 7.1: Normalised contingency table comparing RISP and pathologist A's segmentation masks.

		Pathologist B	
		T	N
RISP (B)	T	**0.216**	0.095
	N	0.097	**0.593**

TABLE 7.2: Normalised contingency table comparing RISP and pathologist B's segmentation masks.

Comparison	**Type 1**	**Type 2**	**Type 3**
RISP (A), Pathologist A	0.291 ($\pm$0.097)	**0.604** ($\pm$0.161)	0.107 ($\pm$0.117)
RISP (B), Pathologist B	**0.305** ($\pm$0.119)	0.572 ($\pm$0.202)	0.123 ($\pm$0.122)
Pathologist A, Pathologist B	0.227 ($\pm$0.144)	0.593 ($\pm$0.218)	**0.180** ($\pm$0.227)
RISP (A), Pathologist B	0.297 ($\pm$0.107)	0.563 ($\pm$0.200)	0.141 ($\pm$0.140)
RISP (B), Pathologist A	0.294 ($\pm$0.107)	0.617 ($\pm$0.153)	0.089 ($\pm$0.099)

TABLE 7.3: Proportions of Type 1, Type 2 and Type 3 disagreements between RISP and manually-obtained segmentation masks.

Automated segmentation masks resulted in a higher proportion of Type 1 disagreements which were previously reported to have minor impact on IHC scores (Section 4.3.1). This suggests a large proportion of disagreements can be disregarded as minor discrepancies with little or no impact on IHC assessment. As shown in Figure 7.1, a higher proportion of Type 2 disagreements were observed between manual and automated segmentation masks. Proportion of remaining Type 3 disagreements closely aligned with Type 3 disagreements between pathologists.

Examples of disagreement images for one TMA spot are shown in Figure 7.2. Here, disagreements between manual segmentation masks appear to be less noisy than disagreements between automated and manual segmentation masks. As manual annotations are not accurate to the pixel level and automated analysis is performed on a superpixel by superpixel basis, this result is not surprising. In the example shown

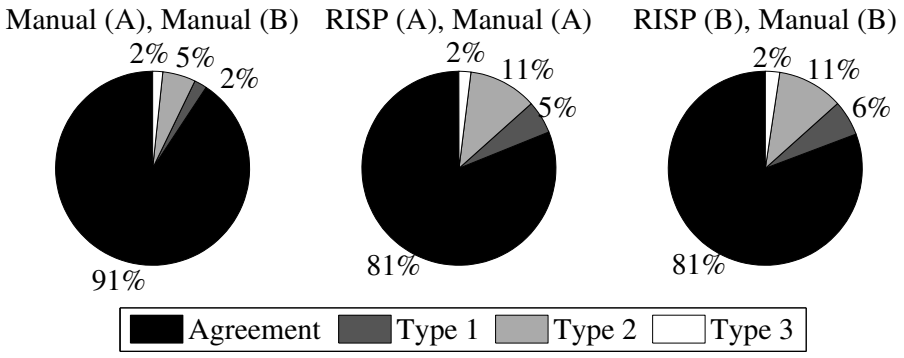


FIGURE 7.1: Pie charts showing distribution of agreements, and Type 1, Type 2 and Type 3 disagreements between manual and RISP segmentations.

in Figure 7.2, pathologist B's segmentation mask was more coarse than pathologist A's segmentation mask. As a result, a higher proportion of Type 2 disagreements are noticeable between RISP and pathologist B's segmentation mask. However in most examples, hand-drawn annotations showed strong agreements and therefore disagreement images were similar regardless of who was used to train the system. When comparisons were reversed such that automated segmentations were compared with manual segmentations from the pathologist *not* used to train the system (Table 7.3), proportion of disagreement types were similar as were κ agreements at the pixel level: agreements between RISP (A) and pathologist B were $\kappa = 0.807$, and between RISP (B) and pathologist A were $\kappa = 0.808$.

7.3.2 IHC scoring

Intensity scores and percentage of positive cells were measured by the Aperio IHC Nuclear algorithm when provided with segmented tumour regions. A Bland Altman plot of percentage of positive ER cells is shown in Figure 7.3. Standard deviations were around 20% with most TMAs showing strong agreements between positive cells identified in automated and manual segmentations. Large disagreements are noticeable on the left hand-side of the Bland Altman plot. Here, cells were underestimated in TMAs where there were higher proportions of positive cells.

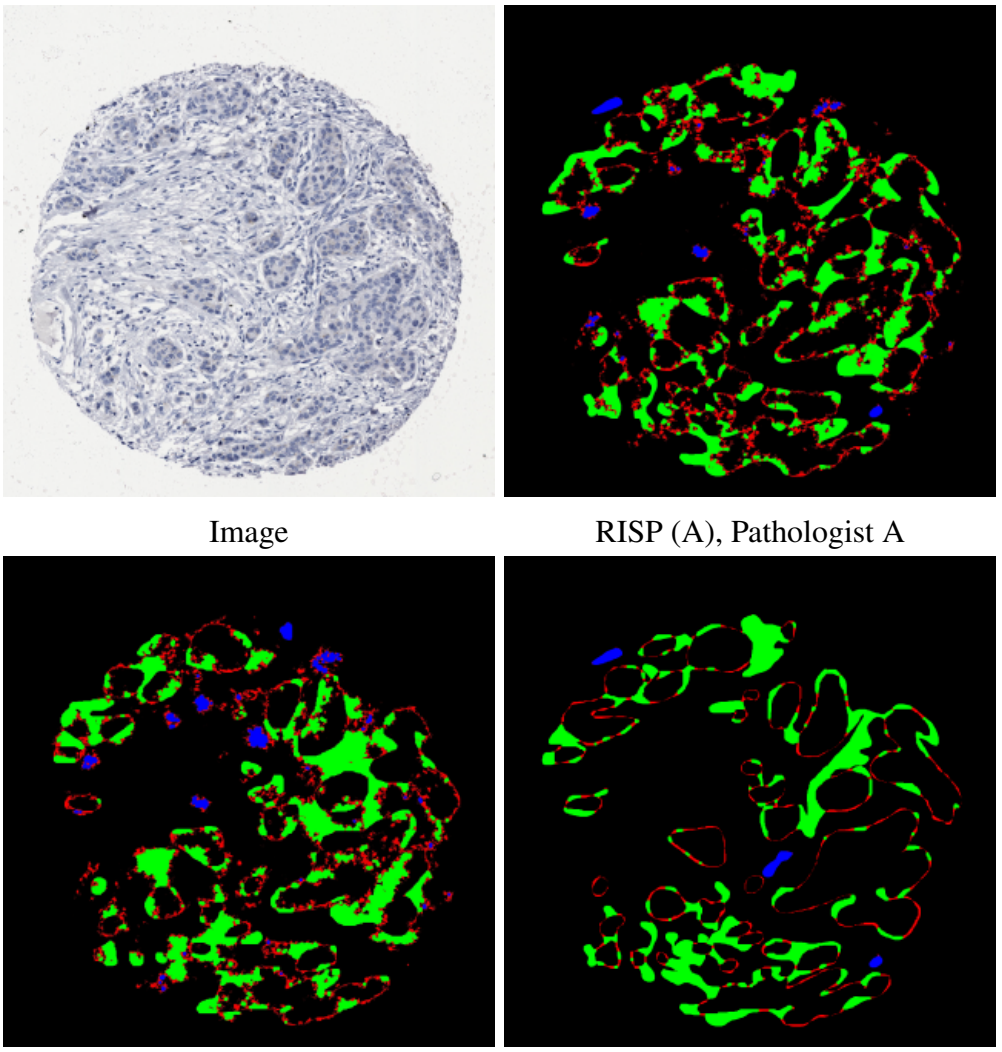


FIGURE 7.2: Original image (left) and disagreement images comparing automated and manual segmentation masks from two pathologists (pathologist A, pathologist B). Disagreements are shown for Type 1 (red), Type 2 (green) and Type 3 (blue).

Agreements between scores calculated in Aperio were reported separately for intensity, and Allred and Quickscore proportion scores in terms of a two-rater weighted Kappa-squared statistic, $\hat{\kappa}$ [33] (Table 7.4). Inter-pathologist agreement was 0.957 for intensity scores, and 0.969 and 0.987 for proportion scores for Allred and Quickscore, respectively. In comparison, automated segmentations on average produced agreements of 0.893 for intensity scores, and 0.848 (Allred) and 0.877 (Quickscores) for proportion scores. Intensity scores revealed strong agreements between automated and manual segmentation masks, approaching inter-rater agreements between experts. Furthermore, intensity scores were consistent regardless of who trained the system.

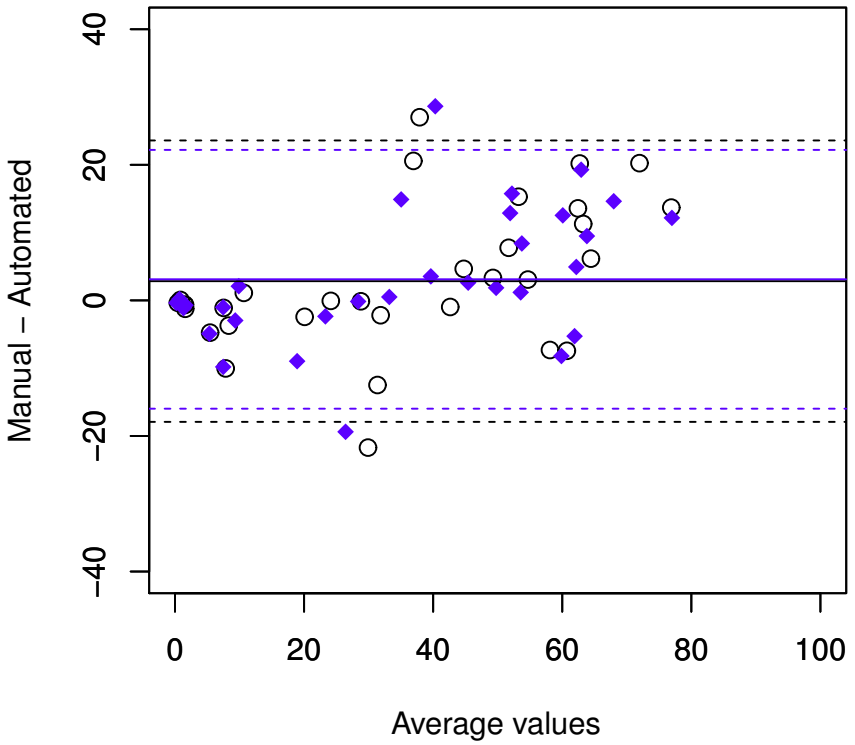


FIGURE 7.3: Bland Altman plot of percentage of positive cells identified in Aperio. TMA spots are shown by black dots (pathologist A) and blue diamonds (pathologist B).

Agreements between proportion scores were slightly lower, suggesting more work can be done to improve precision of cell boundaries.

Agreement for total Allred scores and Quickscores were computed by summing intensity and proportion scores (Table 7.5). Comparisons between automated and manual segmentation masks resulted in average $\hat{\kappa}$ agreements of 0.911 (Allred) and 0.922 (Quickscore). Reported agreements approach inter-rater agreements of 0.980 and 0.989 and show the potential of using automation to generate IHC measurements, achieving similar IHC scores retrieved from manual tumour segementation masks. IHC scores for all 32 TMAs are shown in Figure 7.4 in the form of a histogram plot.

	Intensity		**Proportion**			
			Allred		Quickscore	
	Manual (A)	Manual (B)	Manual (A)	Manual (B)	Manual (A)	Manual (B)
RISP (A)	0.915	0.871	0.848	0.858	0.872	0.868
RISP (B)	0.915	0.871	0.839	0.848	0.885	0.881
Manual (A)	-	0.957	-	0.969	-	0.987

TABLE 7.4: $\hat{\kappa}$ agreements for intensity and proportion scores computed from measurements obtained from the Aperio IHC algorithm.

	Allred		**Quickscore**	
	Manual (A)	Manual (B)	Manual (A)	Manual (B)
RISP (A)	0.913	0.913	0.921	0.916
RISP (B)	0.908	0.909	0.929	0.923
Manual (A)	-	0.980	-	0.989

TABLE 7.5: $\hat{\kappa}$ agreements for computed Allred scores and Quickscores.

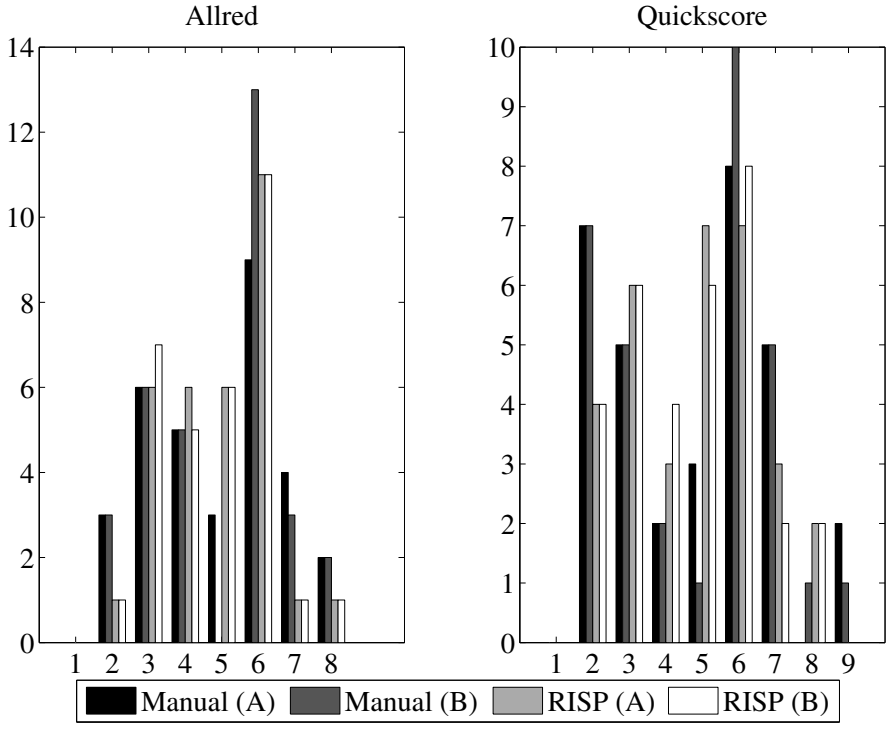


FIGURE 7.4: Histogram plot of Allred scores and Quickscores extracted from manual and automated segmentation masks.

7.3.3 ER treatment

The dichotomy of tumour into ER+ve and ER-ve is essential for treatment decisions for endocrine therapy in clinical practice. To determine the impact of computed IHC score for such treatment decision-making, TMA spots were labelled as ER+ve or ER-ve using commonly used cut-offs implemented in clinical practice.

The Allred cut-off mark (> 2) [5], resulted in almost complete agreement between all reported segmentations, with the exception of a single TMA spot which was labelled ER-ve using automated segmentation masks; as such this would have resulted in *no* treatment for the patient. Using Quickscores (cut-off > 3) [36], two TMA spots were labelled as ER+ve from automated segmentations and the same spots labelled ER-ve from manually obtained segmentations. The remaining 30 spots, 20 ER+ve and 10 ER-ve, were in complete agreement across all segmentations.

Whilst overall there was strong agreement in treatment decisions, results varied between scoring systems and non-standardised cut-offs. These discrepancies suggest more work may be required before automation can be used unreservedly for treatment decisions as suggested for studies comparing visual and automated assessment of Ki-67 markers in breast cancers [46, 98].

7.4 Summary

In this chapter, automated segmentation masks were compared to manually annotated tumour regions for the purpose of IHC scoring. Clinical evaluation of automated tumour localisation revealed on average around 30% of pixel disagreements in segmentation masks relate to minor misalignment of drawn tumour boundaries, termed Type 1 disagreements. Tumour classification differences between automated and manual segmentation masks rarely resulted in a change of IHC score (Allred: $\hat{\kappa} = 0.911$; Quickscore: $\hat{\kappa} = 0.922$). Using the exemplar of nuclear ER staining, the use of automation proposed in this thesis hold promise for reducing the expert pathology time

required and speeding up analysis of IHC stained TMAs from large data sets drawn from clinical trials. The potential usage of automation in clinical practice is discussed further in the following chapter.

Chapter 8

Discussion and Conclusions

With the latest advancements in digital pathology, enabling rapid collection of digital slides, image analysis plays a key role in the day-to-day role of a pathologist. In cancer research, protein expression analysis enables understanding about the diagnosis, treatment and development of tumours at the cellular level. However integrating expert pathology knowledge in an automated system is a challenging task due to complex cellular structures and patterns, and high variability within tissue samples.

Immunohistochemistry (IHC) is important for understanding the distribution of biomarkers such as ER thus supporting new discoveries for diagnosis, prognosis and treatment of cancer. The main hurdle in IHC assessment in digital pathology is the localisation of tumours which is currently performed by manually tracing tumour boundaries. An automated solution can significantly improve throughput and potentially refocus pathologists' workloads.

In the past, tumour localisation has been approached as a pixel-level segmentation problem (Chapter 3). However when observed manually, texture of tissue at the microscopic level is (a) arbitrarily oriented and therefore unsuitable for some state-of-the-art texture features, (b) comprises of complex structural information at multiple scales i.e. close inspection of tissue shows detail at the cellular level whereas ductal/lobular

structures are better observed at lower magnifications, and (c) contains within it complex relationships between tissue structures which are important for distinguishing healthy from abnormal tissue.

8.1 Rotation Invariant Superpixel Pyramid

To capture important structural information in histopathology images, superpixels were adopted (Chapter 6). Superpixels provide rich descriptive information about cellular structures in histopathology. For example, stromal regions tend to encompass elongated superpixels whereas in regions containing lumen and fat, more compact superpixels are generated. In a novel technique called the Rotation Invariant Superpixel Pyramid (RISP), superpixel properties were encoded in superpixel visual words. A rotation invariant pyramid representation was adopted to incorporate information about spatial configuration of superpixels.

RISP showed considerable improvement compared to a similar method reported by Gorelick *et al.* [56] which also incorporated superpixels in a rotation invariant manner. However, instead of using pixel-level features to encode surroundings, superpixels were adopted, which were shown to be more effective for tumour localisation. Results also showed that classification of superpixel features without context was not as effective as pixel-level classification (Figure 6.8), suggesting that spatial information captured from surrounding superpixels is key for localising tumour. Without this information performance was poor. Furthermore, increasing the number of annuli at a single scale resulted in minor improvement. Moreover, further gain in performance was achieved by capturing superpixels at multiple scales.

A theoretical comparison of image features adopted in spin-context and RISP reveals time complexity was significantly reduced by adopting a superpixel representation. The complexity of 3600 x 3600 pixels was reduced to $50,000$ superpixels in RISP; reduction by a factor of $\sim$250. To combat high computational costs associated with spin-context, classification was performed on a regular grid (as opposed to per pixel)

whereas in RISP the baseline superpixel representation was retained. The trade-off of the RISP representation is space complexity. Space complexities associated with image features adopted in spin-context and RISP are summarised as follows:

Spin-context: $M\left(2N_DN_I + 3N_H\right)$, where N_D and N_I are the number of distance and intensity bins, respectively, and N_H is the number of differential invariants (including the zeroth order term). Spin intensity features were computed for 2 scales and differential invariants for 3 scales.

RISP: $Z\Big(K\left(A^L - 1\right) + 24\Big)$; where K is the number of codewords. 24 is the number of superpixel features which were appended to the RISP representation.

In reported experiments lengths of feature descriptors for spin-context image features and RISP were $215M$ and $1424Z$, respectively; where Z is the number of superpixels and M is the number of grid points, or pixels, in an image.

In terms of clinical usage, RISP shows potential for various applications. As there are no restrictions to the underlying image representation, RISP is applicable to IHC membrane and cytoplasmic stains, and other cancer types (e.g. lung, prostate). Whilst non-nuclear stains were not considered in this thesis, theoretically RISP can be adapted to other IHC markers, or indeed in a multi-class framework.

As well as being instrumental for clinical research, RISP shows potential to further recent research in 3D reconstruction of histopathology digital slides [20], whereby acquiring manual annotations for serially sectioned tissue is a time-consuming task. Here, RISP can potentially provide large volumes of labelled tumour regions to provide further insight in breast cancer development. In addition, RISP can be extended for 3D application by simply substituting annuli with spherical shells.

Whilst RISP captured essential structural information enabling accurate tumour classification, in practice superpixels were shown to capture additional complex properties in tissue.

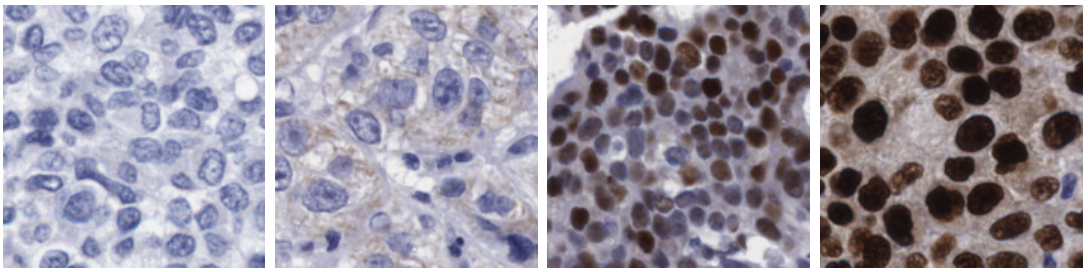


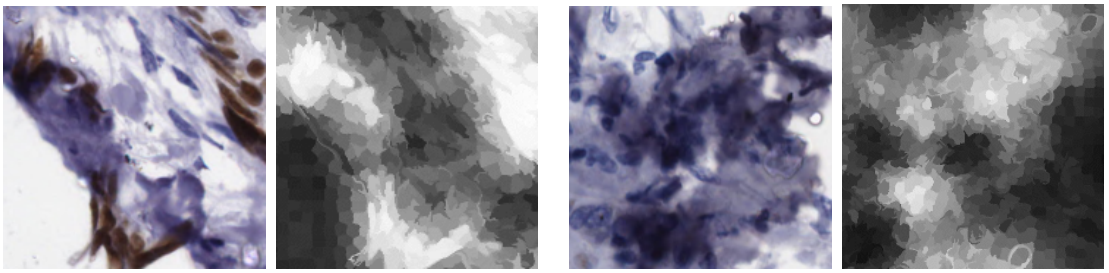
FIGURE 8.1: Image patches in which tumour cells with various IHC staining strengths were correctly classified.

IHC staining: One of the challenges of tumour localisation in ER-stained tissue is distinguishing ER+ve cell nuclei from cancer cell nuclei. Whilst in 80% of breast cancer cases, cancer cells will express ER, it is not always the case. Furthermore, ER+ve cell nuclei which are indeed cancerous, vary in appearance and texture between different staining strengths. Remaining ER-ve cancerous cell nuclei appear similar to healthy epithelial cells and are therefore difficult to distinguish.

This level of complexity at the cellular level was modelled successfully in RISP. As more than one superpixel was used to model cell nuclei, classification of RISPs showed healthy cells were distinguished from cancer cells, regardless of IHC staining strength. Figure 8.1 shows some image patches in which RISP showed successful labelling of cancer cells. Regardless of the staining strength, tumour localisation was strong.

Tissue folding: Artefacts in digital slides come in many forms and are introduced through the various stages involved in preparation of tissue. Most isolated artefacts (bubbles, dirt) posed few problems in the RISP representation, as extraneous superpixels were insignificant in the overall distribution of visual words captured within the circular support window. In level 0 of RISP, noisy superpixels had little impact in the BoS representation whilst at higher levels, spatial information within annuli was still retained.

Tissue folding introduced unexpected complexities in the reported dataset, since in some cases tumour regions were encased within these artefacts (Figure 8.2(a)).



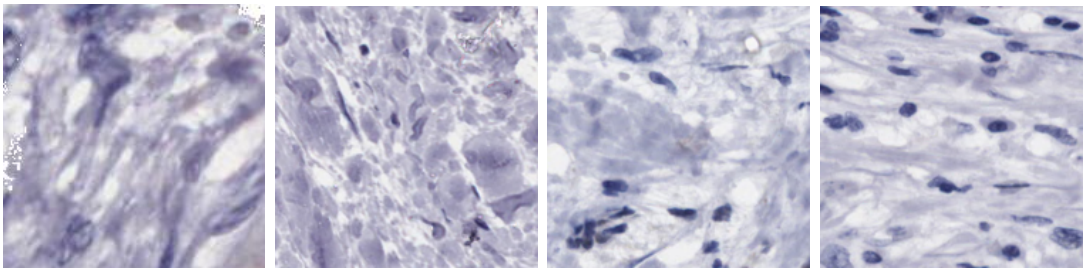
(a) (b)

FIGURE 8.2: Image patches containing folded tissue (left) and superpixel classification maps produced by RISP (right).

When tissue folding occurs, the true nature of the tissue is somewhat altered, commonly referred to as "core loss" [116]. In some cases cell nuclei are still visible. As shown in Figure 8.2(a), in the case of ER+ve stained cells, RISP successfully located cancer cells amongst folded tissue. However misclassification occurred in regions where ER-ve cell nuclei were observed. In Figure 8.2(b), healthy epithelial tissue was incorrectly classified as cancerous due to the irregular structure surrounding ER-ve cell nuclei. Here, superpixels were assigned high tumour probabilities. It is expected that with more examples of artefacts during training, classification can be improved in these areas. Alternatively, there is scope to remove folded tissue as a pre-processing step in RISP. The FDA-approved Aperio software [11] used to extract IHC scores in Chapter 7, embeds a tool for eliminating folded tissue prior to computing cell measurements.

Healthy structures (fat, stroma, lumen): Classification of fat and lumen exhibited few problems in the reported dataset, as these structures are clearly distinguishable from (healthy or cancerous) cell nuclei; furthermore the texture in these areas is uniform. By utilising surrounding superpixels, classification was relatively straightforward.

Identification of stromal regions is a more difficult problem as complex structures such as fibroblasts are encased within stroma. Furthermore, in invasive cancers when stroma combines with cancer cells, tumour is less distinguishable. Figure 8.3(a) shows one of the most difficult cases in the reported dataset.



(a) (b) (c) (d)

FIGURE 8.3: A misclassified stromal region (a) and stromal regions correctly classified as non-tumour ((b)–(d)).

Here, RISP incorrectly labelled stroma as tumour. However there was considerably large regions of tumour surrounding this image patch and therefore tumour proximity, as well as appearance, resulted in poor performance. In cases where stroma did not incorporate other tissue components, classification was better (Figure 8.3(b) - Figure 8.3(d)). In Figure 8.3(d), where there was a high proportion of stromal cells, classification of healthy structures was still accurate.

Encased lumen: An interesting observation in the dataset was the encasement of lumen within tumour regions, as shown in Figure 8.4(a). When presented with this image patch, RISP labelled lumen regions with low tumour probabilities (Figure 8.4(d)). Manual labels acquired from expert pathologists indicated these regions were cancerous (8.4(b), 8.4(c)). Arguably, lumen regions can be labelled as healthy structures captured within tumour regions. As pathologists were not given direction for annotating such regions, both experts instinctively encased lumen within tumour contours. Note, this may be a result of using digital annotation software, as it was not obvious how to draw embedded contours and furthermore this requires additional effort from the operator.

In related work by Chomphuwiset *et al.* [29], lumen encasement was exploited to segment bile ducts in liver tissue. It is therefore debatable as to how manual labels of this kind should be treated during evaluation. Furthermore, there is potential to optimise the annotation procedure to retrieve more accurate labels

(a) (b) (c) (d)

FIGURE 8.4: Image patch containing lumen encased within tumour (a); annotations by expert pathologists (pathologist A (b); pathologist B (c)); and superpixel probabilities generated by RISP (d), where white indicates tumour and black non-tumour.

(see Section 9.3). In the case of IHC scoring, lumen has little impact on resulting scores. However, for other measures such as tumour burden where area of tumour is measured, lumen encasement can be problematic.

To summarise, the usage of superpixels showed a range of benefits for modelling complex patterns and structures in histopathology images. Contextual information from surrounding superpixels in the circular support window was shown to have significant impact on performance, revealing the importance of contextual superpixel surroundings.

8.2 Capturing context from posterior probabilities

Whilst contextual information can refer to a window from which neighbouring textures and patterns are modelled, in this thesis, context extracted from posterior probability maps was also explored. In Chapter 5, tumour probabilities from learned classification maps were captured in the form of a context descriptor. The intuition was that classification of tumour for a particular location is dependent on context from its surroundings. For example, a pixel/superpixel surrounded by tumour labels is also highly likely to incorporate tumour. In Chapter 5, a method for incorporating context was described, called spin-context. This work is an extension of auto-context by Tu and Bai

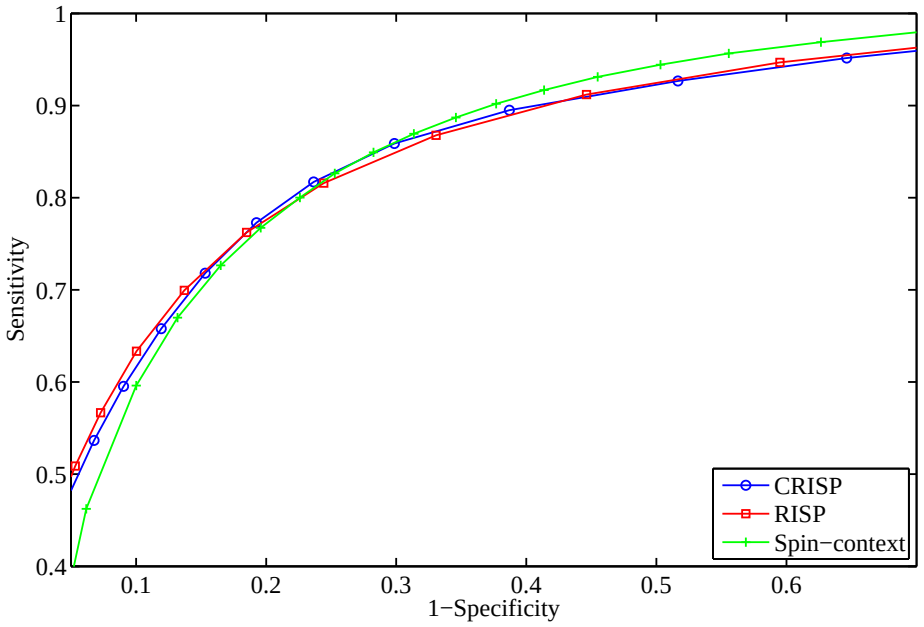


FIGURE 8.5: ROC curves for two iterations of CRISP and boundary sensitive spin-context, and 3-level RISP.

[137] and was modified to provide rotation invariance. Results showed incorporation of context in this manner improved performance.

To extract context information from superpixel classification maps, in Chapter 6, an alternative framework called Contextual RISP (CRISP) was proposed. In CRISP, the RISP representation was altered to model posterior probabilities at multiple scales. Figure 8.5 shows a comparison of CRISP with the boundary sensitive spin-context method proposed in Section 5.5. There was a noticeable improvement between RISP and spin-context at lower sensitivities. However there was little performance gain between RISP (i.e. one iteration of CRISP) and two iterations of CRISP. Results suggested that image-level RISPs already captured contextual information. However there are additional parameters to be explored in the CRISP setup, which is reserved for future work (Section 9.1).

To prevent over-fitting in CRISP, an experimental setup called nested cross-validation (Section 6.5) was proposed to ensure separation of training and validation data across

	Inter-rater	Automated	
		Spin-context	RISP
κ	0.908	0.829	0.811

TABLE 8.1: κ agreements between manually and automatically-obtained segmentation masks. Automated κ agreements are reported as an average of agreements between each automated method trained on pathologist A and pathologist B, and manual segmentation masks.

multiple folds. To achieve this, in each fold the dataset was partitioned into V sub-folds. Due to the small dataset used in reported experiments, there are some concerns about whether there was sufficient training data available to represent highly variable breast TMAs. In reported CRISP experiments, only a few training samples were made available in each sub-fold. In a dataset containing a total of 32 TMA spots, only 14, $\frac{N(U-1)}{UV}$, spots were assigned for training in each sub-fold when $U = 8$ and $V = 2$. As such it is anticipated a larger dataset would give a more representative training set, appropriate for nested cross-validation.

8.3 Clinical impact of automated tumour localisation

In the computer vision literature, tumour image analysis is often evaluated at the pixel-level without evaluating the clinical implications of using automation in practice. To provide some insight into the effects of using automated tumour localisation for IHC assessment, IHC scores were computed and compared between manually-obtained and automatic segmentation masks (Chapter 7). In reported studies, manual segmentations were hand-drawn tumour contours drawn by expert pathologists. To measure inter-rater agreement, manual segmentations were acquired from two pathologists, pathologist A and pathologist B. Pixel-level κ agreements between manual and (spin-context, RISP) automated segmentation masks are shown in Table 8.1. Agreements between computed IHC scores (Allred and Quickscore), and intensity and proportion scores, are reported in Table 8.2. Agreements are shown as an average over pathologist A and pathologist B.

	Manual	Automated
Intensity	0.957	0.893
Allred		
Proportion	0.848	0.848
Total	0.980	0.911
Quickscore		
Proportion	0.877	0.877
Total	0.989	0.922

TABLE 8.2: Overview of $\hat{\kappa}$ agreements for Allred scores and Quickscores computed from automated (RISP) and manual segmentation masks.

Pixel-level inter-rater agreements between expert pathologists revealed strong agreements ($\kappa = 0.908$). As both pathologists recruited for these studies are fully trained with several years of experience, it is anticipated that agreements will differ between laboratories and different levels of expertise. Regardless, by training automated systems proposed in this thesis on labels provided by clinical pathologists, RISP showed good agreements with manual segmentation masks. With further advancements in medical image analysis, agreements between automated and manual tumour segmentation show potential to increase further.

When pixel-level agreements between pathologists were analysed in more detail, it was found that 23% of disagreements correlated to minor disagreements, termed Type 1 (Section 4.3). When automated segmentation masks were analysed in a similar manner, higher proportions of disagreements were found to be of Type 1 (around 30%) suggesting a large proportion of disagreements in automated segmentations are inconsequential for IHC scoring. It is questionable whether pixel-level analysis is appropriate for image analysis systems intended for clinical usage. Instead, the author suggests a deeper understanding of the intended use of the system is required, as hypothesised by Gurcan *et al.* [58]. In this thesis, disagreements were categorised to distinguish minor misalignment of hand-drawn tumour boundaries from disagreements which impacted extracted IHC scores. For other applications, such as tumour grading, an alternative method of categorising disagreements are best considered.

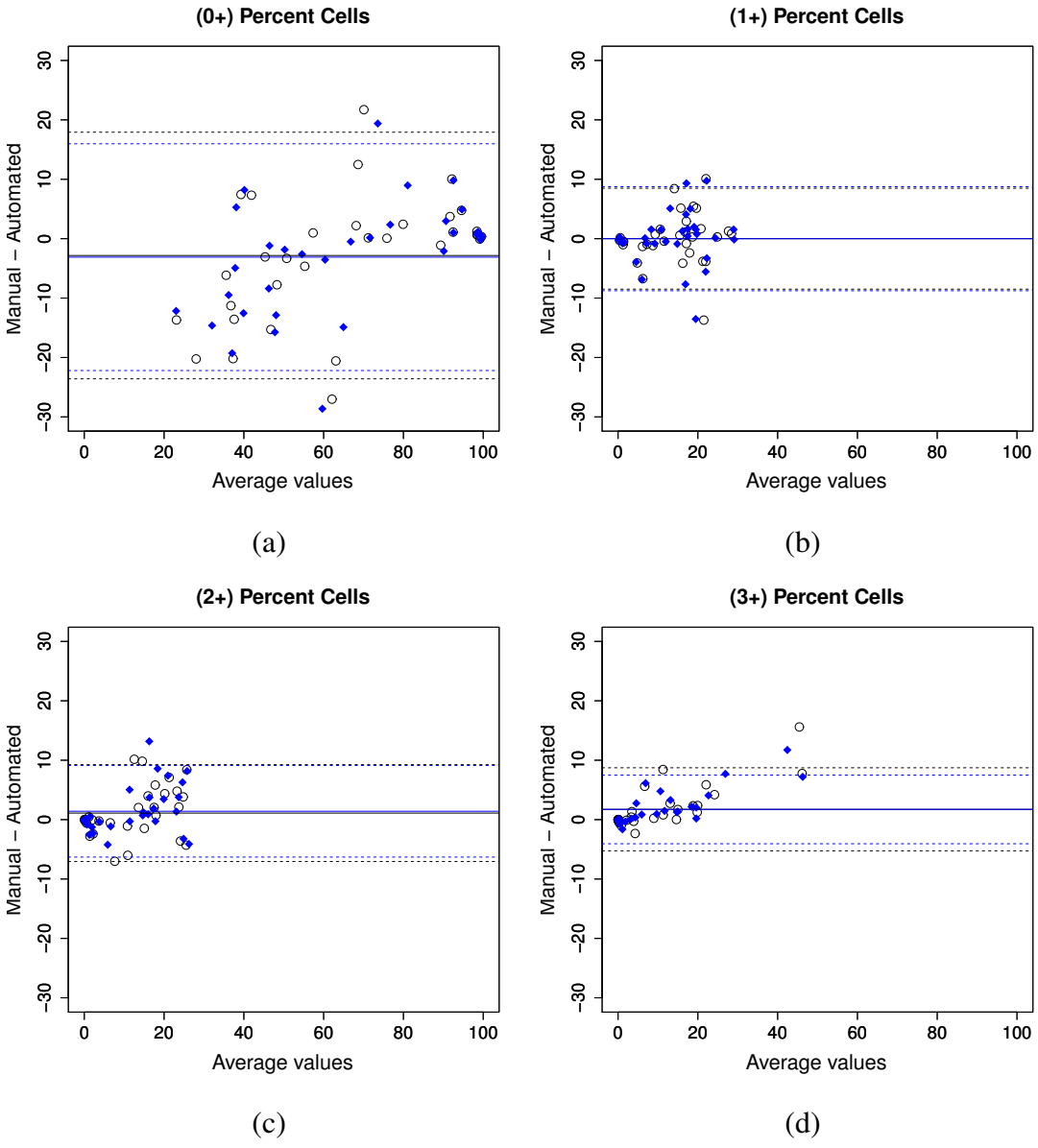


FIGURE 8.6: Bland Altman plots comparing negatively (a), weakly (b), moderately (c) and strongly (d) stained cell nuclei extracted from manual and automated segmentation masks.

Despite lower inter-rater agreements between automated and manually-obtained segmentation masks at the pixel-level, computed IHC scores from automated segmentation masks were in strong agreement with scores computed from manual segmentation masks (Allred; $\hat{\kappa} = 0.911$, Quickscore; $\hat{\kappa} = 0.922$). When agreements were evaluated between intensity and proportion scores, automation approached inter-rater agreements between experts. However agreements between proportion scores extracted from automated and manual segmentation masks were noticeably higher. Standard deviations indicated that there were $\pm 20\%$ disagreements between manual and automated segmentation masks, regardless of who was used to train the system. Figure 8.6 shows Bland Altman plots comparing percentage of negatively (0+), weakly (1+), moderately (2+) and strongly (3+) stained cell nuclei that were identified in the Aperio IHC Nuclear software. Large disagreements occurred amongst cell counts of negatively stained nuclei as shown in Figure 8.6(a). When considering the average percentage of negatively stained cells between manual and automated segmentation masks, percentages were over-estimated when there were fewer negatively stained cells and under-estimated when there were more. Notice the average proportion of negatively stained cell nuclei were rarely close to zero, unlike proportion of positively stained cells. In the majority of TMA spots, percentage of positively (i.e. weakly, moderately and strongly) stained cells rarely exceeded 30%.

Despite lacking agreements between proportion scores, this had little effect on overall IHC scores (i.e. summation of intensity and proportion scores) as shown in Table 8.2. Allred scores and Quickscores extracted from automated segmentation masks closely aligned with inter-rater agreements between scores extracted from manual segmentation masks. Given this outcome, it is anticipated that with a larger number of tumour samples, application of automated annotations will conclude similar outcomes to more labour intensive manual annotations. Thus, the benefits of automation extend beyond the reproducibility of IHC scores to include changing the focus of research pathologists' workloads.

Using the exemplar of nuclear ER staining, methods of automated tumour image analysis employed in this thesis hold promise for reducing the expert pathology time required and speeding up analysis of IHC stained TMAs from large data sets drawn from clinical trials. Automation produced positive outcomes, approaching inter-rater agreements between experts in pathology. For the time being automation may not be used unreservedly for treatment decision-making (Section 7.3.3), but may be applicable to large clinical studies where availability of manual skilled pathologists is lacking.

8.4 Contributions

The work presented in this thesis shows automated tumour localisation can be performed reliably and accurately, when compared to benchmark performance measured between expert pathologists. The main contributions in this thesis are summarised as follows.

1. A method called spin-context was described in which context information was extracted from learned classification maps which was shown to improve performance in an iterative framework. Results showed spin-context surpassed auto-context at lower sensitivities, matching performance at higher sensitivities. An extension to spin-context was proposed to remove background interference by excluding context locations outwith TMA spot boundaries. Results showed this approach, boundary-sensitive spin-context, improved performance in all spin-context iterations.

2. To capture essential structural information in tissue, a Rotation Invariant Superpixel Pyramid (RISP) representation was proposed. In RISP, frequencies of superpixel visual words and spatial configuration of superpixels were captured at multiple scales in a pyramid structure. In each pyramid level of RISP, a spatial Bag-of-Superpixels (S-BoS) was proposed to capture spatial information in the

form of equally-spaced annuli. An experiment was performed to compare Bag-of-Superpixels (BoS), Spatial Bag-of-Superpixels (S-BoS) and RISP, of which RISP was superior.

3. Spin-context was adapted to incorporate classified superpixels. Context-level RISPs were proposed in which posterior probabilities were modelled in a RISP form. In each level of the context-level RISP, tumour distributions were captured within equally-spaced annuli. Image-level and context-level RISPs were combined in a novel framework called Contextual RISP (CRISP). In CRISP, superpixel classification maps were iteratively updated and used to construct context-level RISPs. Compared to the original RISP representation, CRISP showed comparable performance.

4. Tumour localisation was reviewed and evaluated for clinical assessment, specifically IHC scoring. A study was designed to measure the impact of utilising automated RISP tumour segmentations to compute IHC scores from ER-stained TMAs. Inter-rater agreements of $\kappa = 0.908$ were found between manual tumour segmentations drawn by two expert pathologists. A comparison between automated and manual segmentation masks revealed automation now approaches inter-rater agreements (on average $\kappa = 0.811$). Extracted IHC scores were then compared between manual and automated segmentation masks. Results showed IHC scores computed from automated segmentations revealed strong agreements with scores extracted from manual segmentation masks (Allred: $\hat{\kappa} = 0.911$; Quickscore: $\hat{\kappa} = 0.922$), approaching inter-rater agreements between experts (Allred: $\hat{\kappa} = 0.980$; Quickscore: $\hat{\kappa} = 0.989$).

   Pixel-level disagreements between tumour segmentation masks were categorised into three types: Type 1, Type 2 and Type 3 disagreements. Type 1 disagreements were found to correlate to minor discrepancies between hand-drawn tumour boundaries with little effect on extracted IHC scores.

To conclude, automated IHC assessment shows potential to further molecular analysis of protein expression in cancer research. Methods for automated tumour localisation

described in this thesis showed similar outcomes to experts in pathology and thus hold promise for improving clinical workflow.

Chapter 9

Recommendations

The following recommendations are offered for future directions in this research.

9.1 Exploring CRISP parameters

Evaluation of CRISP (Section 6.4) revealed that the framework offered little gain in tumour classification accuracy compared to RISP. However there are a range of parameters that are yet to be explored, some of which are:

- The number of levels, L, in context-level RISPs.
- The number of bins, B, used to model posterior tumour distributions in context-level RISPs.
- Varying widths of annuli in the circular support window, such that numbers of superpixels are equally distributed between each row of the context-level RISP.
- Alternatively, incorporating an energy function to place higher/lower emphasis on superpixel nearby or further away from the central superpixel in the circular context support window.

Future work will investigate the impact of varying these properties. It is anticipated an alternative setup in CRISP will improve upon performance achieved using image-level RISPs alone.

To evaluate CRISP, a nested cross-validation setup (Section 6.5) was adopted which partitions folds to ensure separation of training and validation sets. Repeatedly partitioning the dataset meant fewer samples were made available for training in sub-folds. Given that the dataset used in reported experiments was lacking in annotated data, expanding the dataset will be beneficial for future research. In addition to gathering more data, future work will investigate how the number of training samples in each sub-fold in nested cross-validation, impact overall performance. This will help determine the optimal number of annotated TMA spots required to ensure a representative training set.

9.2 Contextual superpixel factor graph

In context-level RISPs, direct relationships between neighbouring superpixels are not well captured. Given that a superpixel representation is a non-uniform structure, accurate information about relative distances and relationships between superpixels can be important for modelling complex patterns. To encompass this information a graphical representation shows potential benefits, whereby superpixel posterior probabilities are represented as nodes.

Factor graphs enable probability distributions to be derived from graphical models. In computer vision, a common usage of factor graphs is "message passing" which enables inference to be performed as a joint likelihood across a patch or image. Future work will investigate application of factor graphs to superpixel representations whereby posterior probabilities can be captured across an entire TMA spot, with higher emphasis placed on direct superpixel neighbours. Unlike in CRISP and spin-context, this will enable context to be captured globally which can potentially complement information captured in context-level RISPs. Chen *et al.* [28] proposed an

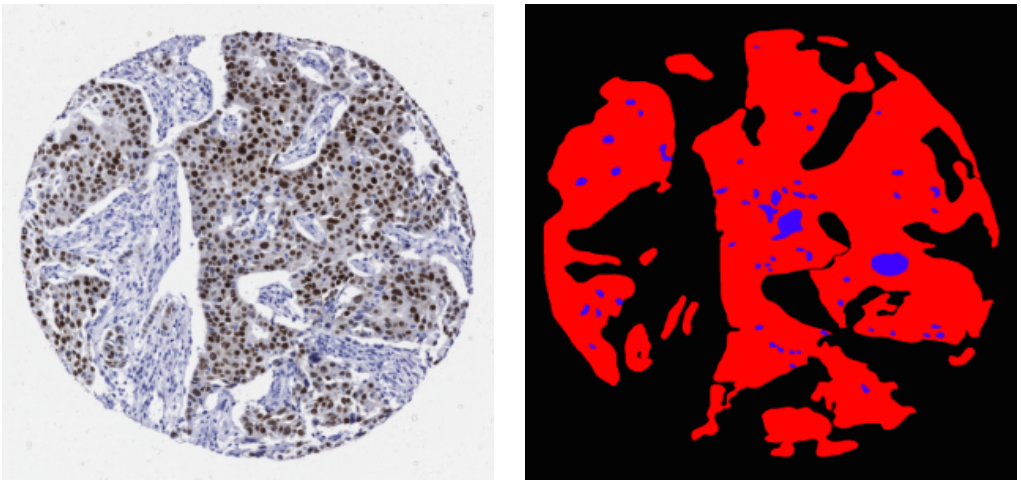


FIGURE 9.1: TMA spot (left) and annotation (right) drawn by an expert pathologist. Annotation labels are shown for invasive tumour (red) and encased healthy structures (blue).

efficient manner for performing inference in a graph structure by employing decomposable k-way potentials. This method was developed to model relationships between cellular structures but may also be applicable to a superpixel representation in which the number of neighbours can vary per superpixel. Furthermore superpixel properties can be encoded within the potential function to enable appropriate weight assignments based on superpixel shape, appearance, geometry and location.

9.3 Gathering manual annotations

Figure 9.1 shows a manual annotation of invasive tumour regions (red) which were annotated in a similar manner to segmentation masks reported in this thesis. Whilst annotating this TMA spot, the pathologist also highlighted regions which are not cancerous but are encased within tumour, shown in blue. In this thesis, encased healthy structures such as lumen were not explicitly labelled to ease the annotation procedure. However, the effort to produce additional class labels could potentially provide a more accurate learning system by refining classification function boundaries. By explicitly locating healthy structures encased within "tumour" regions, the quality of collected

annotations can potentially outweigh the benefits of providing large volumes of coarse annotations. This is yet to be investigated.

Acquisition of additional labels, particularly examples which were considered infrequent in the reported dataset (i.e. tissue folding) will also be reserved for future work. It is anticipated with a more representative training set reflecting large variations in the dataset, classification accuracy can be improved further.

9.4 Standardisation across laboratories

Between laboratories, tissue preparative phases e.g. staining conditions, are likely to differ thus introducing differences between tissue sections. Furthermore, as tissue degrades over time, the timeframe between tissue preparation and digital scanning can introduce further complexities. Even scanner specifications (manufacturer, focusing, white balancing etc.) can cause variations to occur between datasets [2]. Whilst in this thesis, histopathology images were drawn from multiple TMAs, images prepared across laboratories is yet to be explored. In future work, comparisons will be performed between tissue samples acquired from multiple laboratories to assess how variations resulting from tissue preparation and acquisition can effect IHC analysis. Similar to evaluation techniques described in this thesis, the impact of automation in clinical practice will be assessed for standardisation.

Related work in stain normalisation has also shown to be beneficial for standardisation across datasets and laboratories [58]. Khan *et al.* [76] showed improvement in tumour segmentation accuracy by using stain normalisation as a preprocessing step. In future work, integration of stain normalisation with methods described in this thesis will be explored, with the potential to improve tumour localisation further.

Appendix A

Superpixel Autocorrelogram

Bag-of-Words (BoW) was originally proposed for information retrieval as a means of representing text (i.e. a sentence) as a bag of individual words. Later it was adopted in the computer vision literature to capture frequency of learned visual words in the form of a one-dimensional histogram [131]. In BoW, visual words are learned using a feature encoding technique such as K-means clustering.

Whilst BoW is a simple yet powerful representation, its main drawback is lack of spatial information. An alternative approach is the codebook correlogram, proposed by Zheng *et al.* [158], which retains spatial information by including the distance distribution of the position of visual words within a codebook dictionary. Zheng *et al.* described a histogram representation with three dimensions: two dimensions indexed visual words and the third equally-spaced spatial distributions. Spatial distributions of "neighbouring" codewords were incorporating in the correlogram, such that a pair of codewords (v_i, v_j) positioned d_{ij} from each other contributed towards the bin (v_i, v_j, d_{ij}) in the three-dimensional correlogram. Figure A.1 illustrates how the codebook correlogram differs from traditional BoW.

However one of the main drawbacks of the correlogram is that it requires $O(D^2Q)$ space, where Q is the number of rings at the current level and D is the dictionary size. In order to improve computational costs by reducing the dimensionality of the correlogram, Huang *et al.* [64] proposed the colour autocorrelogram. In the autocorrelogram

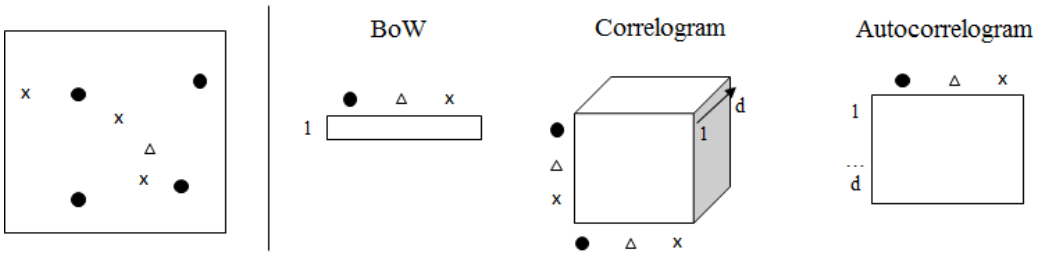


FIGURE A.1: Illustrative comparison of BoW, correlogram and autocorrelogram, given the support window shown on the left.

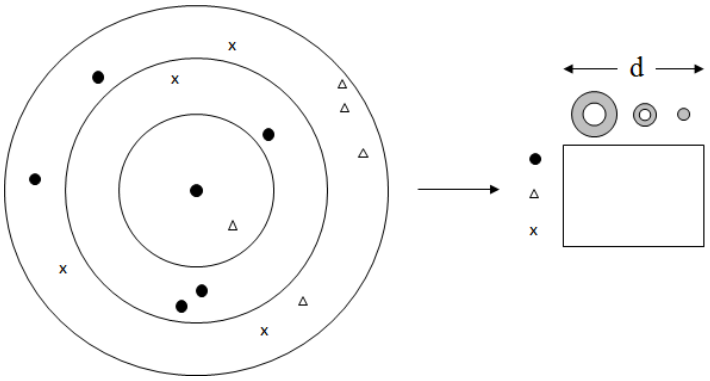


FIGURE A.2: Two-dimensional superpixel autocorrelogram, indexed by visual words and spatial distribution (denoted by shaded regions of the circular support window).

only identical visual words are recorded in the two-dimensional histogram thereby eliminating one dimension from the correlogram. This reduces space complexity to $O(DQ)$ whilst time complexity remains as $O(D^2Q)$. Zheng *et al.* [158] also propose a similar extension to the codebook correlogram, called the self-correlogram.

In the work reported, a *superpixel autocorrelogram* is proposed which captures spatial information between pairs of superpixels. Unlike in the method proposed by Zheng *et al.* , distance is measured in the image space rather than the feature space, thereby encoding relative positioning of superpixel pairs. Similarly to [64], only identical pairs of codewords are counted. However in the superpixel autocorrelogram, circular windows are used instead of a regular grid to retain rotation invariance. As in RISP (Chapter 6), rings are equally-spaced apart and centred on a superpixel to be classified. The distance distribution between *all* superpixel centre points which lie within each ring is captured, resulting in a two-dimensional superpixel autocorrelogram (Figure A.2). In reported experiments (Chapter 6), five equally-spaced annuli were used in the circular support window with 200 superpixel visual words.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 34, pp. 2274–2282, 2012.

[2] S. Al-Janabi, A. Huisman, and P. J. Van Diest, "Digital pathology: current status and future perspectives," *Histopathology*, vol. 61, pp. 1–9, 2012.

[3] G. Alexe, G. S. Dalgin, D. Scanfeld, P. Tamayo, J. P. Mesirov, C. DeLisi, L. Harris, N. Barnard, M. Martel, A. J. Levine, S. Ganesan, and G. Bhanot, "High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates," *Cancer Research*, vol. 67, 2007.

[4] S. Ali, R. Veltri, J. I. Epstein, C. Christudass, and A. Madabhushi, "Adaptive energy selective active contour with shape priors for nuclear segmentation and gleason grading of prostate cancer," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 14, pp. 661–669, 2011.

[5] D. C. Allred, M. A. Bustamante, and C. O. Daniel, "Immunocytochemical analysis of estrogen receptors in human breast carcinomas: Evaluation of 130 cases and review of the literature regarding concordance with biochemical assay and clinical relevance," *Archives of Surgery*, vol. 125, pp. 107–113, 1990.

[6] D. Altunbay, C. Cigir, C. Sokmensuer, and C. Gunduz-Demir, "Color graphs for automated cancer diagnosis and grading," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 665–674, 2010.

[7] T. Amaral, "Analysis of breast tissue microarray spots," Ph.D. dissertation, University of Dundee, 2010.

[8] T. Amaral, S. J. McKenna, K. Robertson, and A. Thompson, "Classification and immunohistochemical scoring of breast tissue microarray spots," *IEEE Transaction on Biomedical Engineering*, vol. 60, pp. 2806–2814, 2013.

[9] American Cancer Society. (2013, August) What are the risk factors for breast cancer? [Online]. Available: http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-risk-factors

[10] Aperio. (2013, August) Aperio PRECISION. [Online]. Available: http://www.aperio.com/lifescience/precision

[11] ——. (2014, December) IHC Nuclear Image Analysis. [Online]. Available: http://tmalab.jhmi.edu/aperiou/userguides/IHC_Nuclear.pdf

[12] M. Arif and N. Rajpoot, "Classification of potential nuclei in prostate histology images using shape manifold learning," in *International Conference on Machine Vision (ICMV)*, Islamabad, 2007, pp. 113–118.

[13] K. Arihiro, S. Umemura, M. Kurosumi, T. Moriya, T. Oyama, H. Yamashita, Y. Umekita, Y. Komoike, C. Shimizu, H. Fukushima, H. Kajiwara, and F. Akiyama, "Comparison of evaluations for hormone receptors in breast carcinoma using two manual and three automated immunohistochemical assays," *American Journal of Clinical Pathology*, vol. 127, pp. 356–365, 2007.

[14] S. Avninder, K. Ylaya, and S. M. Hewitt, "Tissue microarray: A simple technology that has revolutionized research in pathology," *Journal of Postgraduate Medicine*, vol. 54, pp. 158–162, 2008.

[15] A. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot, and A. Madabhushi, "Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 642–653, 2010.

[16] A. Basavanhally, S. Ganesan, M. Feldman, N. Shih, and C. Mies, "Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides," *IEEE Transactions on Biomedical Engineering*, vol. 60, pp. 2089–2099, 2013.

[17] H. Batiffora, "The multitumor (sausage) tissue block: novel method for immunohistochemical antibody testing," *Laboratory Investigation*, vol. 55, pp. 244–248, 1986.

[18] A. H. Beck, A. R. Sangoi, S. L. Leung, R. J. Marinelli, T. O. Neilson, M. J. Vijver, R. B. West, M. Rijn, and D. Koller, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science Translational Medicine*, vol. 3, 2011.

[19] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, pp. 509–522, 2002.

[20] M. E. Booth, D. Treanor, N. Roberts, D. R. Magee, V. Speirs, and A. M. Hanby, "Three-dimensional reconstruction of ductal carcinoma in-situ with virtual slides," *Histopathology*, vol. 66, 2015.

[21] M. Braun, R. Kirsten, N. J. Rupp, H. Moch, F. Fend, N. Wernert, G. Kristiansen, and S. Perner, "Quantification of protein expression in cells and cellular subcompartments on immunohistochemical sections using a computer supported image analysis system," *Histology and Histopathology*, vol. 28, pp. 605–610, 2013.

[22] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, pp. 532–540, 1983.

[23] R. L. Camp, L. A. Charette, and D. L. Rimm, "Validation of tissue microarray technology in breast carcinoma," *Laboratory Investigation*, vol. 80, pp. 1943–1949, 2000.

[24] Cancer Research UK. (2014, May) Breast Cancer. [Online]. Available: http://publications.cancerresearchuk.org/downloads/Product/CS_KF_BREAST.pdf

[25] J. D. Cass, S. Varma, A. G. Day, W. Sangrar, A. B. Rajput, L. H. Raptis, J. Squire, Y. Madarnas, S. K. SenGupta, and B. E. Elliott, "Automated quantitative analysis of p53, cyclin d1, ki67 and perk expression in breast carcinoma does not differ from expert pathologist scoring and correlates with clinicopathological characteristics," *Cancers*, vol. 4, pp. 725–742, 2012.

[26] H. Chang, A. Borowsky, P. Spellman, and B. Parvin, "Classification of tumor histology via morphometric context," in *Computer Vision and Pattern Recognition (CVPR)*, Portland, U.S., 2013.

[27] H. Chang, N. Nayak, P. Spellman, and B. Parvin, "Characterization of tissue histopathology via predictive sparse decomposition and spatial pyramid matching," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 16, pp. 91–98, 2013.

[28] S.-C. Chen, G. J. Gordon, and R. F. Murphy, "Graphical models for structured classification, with an application to interpreting images of protein subcellular location patterns," *Journal of Machine Learning Research*, pp. 651–682, 2008.

[29] P. Chomphuwiset, D. Magee, R. Boyle, and D. Treanor, "Nucleus classification and bile duct detection in liver histology," in *MICCAI Workshop on Machine Learning in Medical Imaging*, Beijing, China, 2010.

[30] ——, "Context-based classification of cell nuclei and tissue regions in liver histopathology," in *Medical Image Understanding and Analysis (MIUA)*, London, U.K., 2011.

[31] K. R. Choudhury, K. J. Yagle, P. E. Swanson, K. A. Krohn, and J. G. Rajendran, "A robust automated measure of average antibody staining in immunohistochemistry images," *Journal of Histochemisty and Cytochemistry*, vol. 58, pp. 95–107, 2010.

[32] D. Cireşan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 8150, pp. 411–418, 2013.

[33] J. Cohen, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, pp. 213–220, 1968.

[34] A. Coons, H. Creech, and R. Jones, "Immunological properties of an antibody containing a fluorescent group," *Experimental Biology and Medicine*, vol. 47, pp. 200–202, 1941.

[35] Definiens. (2014, November) Definiens Tissue Studio. [Online]. Available: http://tissuestudio.definiens.com/

[36] S. Detre, G. S. Jotti, and M. Dowsett, "A "quickscore" method for immunohistochemical semiquantitation: validation for oestrogen receptor in breast carcinomas," *Journal of Clinical Pathology*, vol. 48, pp. 876–878, 1995.

[37] S. J. Dickinson, A. Levinshtein, and C. Sminchisescu, "Perceptual grouping with superpixels," *Pattern Recognition*, vol. 7329, pp. 13–22, 2012.

[38] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, pp. 31–71, 1997.

[39] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Paris, France, 2008, pp. 496–499.

[40] S. Doyle, M. Feldman, J. Tomaszewski, N. Shih, and A. Madabhushi, "Cascaded multi-class pairwise classifier (CascaMPa) for normal, cancerous, and

cancer confounder classes in prostate histology," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Chicago, U.S., 2011, pp. 715–718.

[41] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "Automated grading of prostate cancer using architectural and textural image features," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Virginia, U.S., 2007, pp. 1284–1287.

[42] S. Doyle, C. Rodriguez, A. Madabhushi, J. Tomaszeweski, and M. Feldman, "Detecting prostatic adenocarcinoma from digitized histology using a multi-scale hierarchical classification approach," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2006, pp. 4759–4762.

[43] C. Drummond and R. C. Holte, "What ROC curves can't do (and cost curves can)," in *Workshop on ROC Analysis in AI*, 2004.

[44] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up." *Histopathology*, vol. 19, no. 5, pp. 403–410, 1991.

[45] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[46] S. Fasanella, E. Leonardi, C. Cantaloni, C. Eccher, I. Bazzanella, D. Aldovini, E. Bragantini, L. Morelli, L. V. Cuorvo, A. Ferro, F. Gasperetti, G. Berlanda, P. Dalla Palma, and M. Barbareschi, "Proliferative activity in human breast cancer: ki-67 automated evaluation and the influence of different ki-67 equivalent antibodies," *Diagnostic Pathology*, vol. 6, 2011.

[47] H. Fatakdawala, J. Xu, A. Basavanhally, G. Bhanot, S. Ganesan, M. Feldman, J. E. Tomaszewski, and A. Madabhushi, "Expectation maximization-driven geodesic active contour with overlap resolution (EMaGACOR): Application to

lymphocyte segmentation on breast cancer histopathology," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1676–1689, 2010.

[48] P. F. Felzenswalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision (IJCV)*, vol. 59, pp. 167–181, 2004.

[49] D. J. Foran, L. Yang, O. Tuzel, W. Chen, J. Hu, T. M. Kurc, R. Ferreira, and J. H. Saltz, "A caGrid-enabled, learning based image segmentation method for histopathology specimens," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Boston, U.S., 2009, pp. 1306–1309.

[50] K. Fukushima, "Neocognitron: A self-organizing neural network for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.

[51] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *IEEE 12th International Conference on Computer Vision (ICCV)*, Kyoto, Japan, 2009, pp. 670–677.

[52] W. Gallagher, E. Rexhepaj, and D. Brennan, "Method and system for image analysis," Patent US 8 116 551, February 14, 2012. [Online]. Available: http://www.google.com/patents/US8116551

[53] J. Gerdes, U. Schwab, H. Lemke, and H. Stein, "Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation," *International Journal of Cancer*, vol. 31, no. 1, pp. 13–20, 2002.

[54] D. F. Gleason, "The veteran's administration cooperative urologic research group: histologic grading and clinical staging of prostatic carcinoma," *Tannenbaum, M. Urologic Pathology: The Prostate.*, pp. 171–198, 1977.

[55] P. Gong, Y. Wang, G. Liu, J. Zhang, and Z. Wang, "New insight into ki67 expression at the invasive front in breast cancer," *PLoS ONE*, vol. 8, 2013.

[56] L. Gorelick, O. Veksler, M. Gaed, J. A. Gómez, M. Moussa, G. Bauman, A. Fenster, and A. D. Ward, "Prostate histopathology: Learning tissue component histograms for cancer detection and classification," *IEEE Transactions on Medical Imaging (TMI)*, vol. 22, pp. 1804–1818, 2013.

[57] S. Gould, J. Rodgers, D. Cohen, G. Ellidan, and D. Koller, "Multi-class segmentation with relative location prior," *International Journal of Computer Vision*, vol. 80, pp. 300–316, 2008.

[58] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[59] M. D. Gustavson, B. Bourke-Martin, D. Reilly, M. Cregger, C. Williams, J. Mayotte, M. Zerkowski, G. Tedeschi, R. Pinard, and J. Christiansen, "Standardization of HER2 immunohistochemistry in breast cancer by automated quantitative analysis," *Archives of Pathology and Laboratory Medicine*, vol. 133, no. 9, pp. 1413–1419, 2009.

[60] Z. Hao, Q. Wang, Y. K. Seong, J.-H. Lee, H. Ren, and J.-Y. Kim, "Combining CRF and multi-hypothesis detection for accurate lesion segmentation in breast sonagrams," *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 7510, pp. 504–511, 2012.

[61] R. Haralick and L. Shapiro, *Computer and Robot Vision*. Addison-Wesley Publishing Company, 1992.

[62] J. M. Harvey, G. M. Clark, and D. C. Allred, "Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer," *Journal of Clinical Oncology*, vol. 17, pp. 1474–1481, 1999.

[63] A. Hefiane, F. Bunyak, and K. Palaniappan, "Fuzzy clustering and active contours for histopathology image segmentation and nuclei detection," in *Advanced Concepts for Intelligent Vision Systems*, 2008, pp. 903–914.

[64] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Juan, Puerto Rico, 1997, pp. 762–768.

[65] Indica Labs. (2014, December) Indica HALO inage analysis. [Online]. Available: http://www.histologix.co.uk/technologies/indica-halo-image-analysis

[66] ——. (2014, December) Tissue Microarray. [Online]. Available: http://indicalab.com/tissue-microarray-tma/

[67] International Academy of Pathology. (2014, December) USCAP. [Online]. Available: http://www.uscap.org

[68] V. Jampani, R. Gadde, and P. V. Gehler, "Efficient facade segmentation using auto-context," in *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*, Hawaii, 2015.

[69] T. Jiang, F. Jurie, and C. Schmid, "Learning shape prior models for object matching," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Florida, U.S., 2009, pp. 848–855.

[70] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[71] E. Jurrus, A. Paiva, S. Watanabe, J. Anderson, B. Jones, R. Whitaker, E. Jorgensen, R. Marc, and T. Tasdizen, "Detection of neuron membranes in electron microscopy images using a serial neural network architecture," *Medical Image Analysis*, vol. 14, no. 6, pp. 770–783, 2010.

[72] H. Kalkan, M. Nap, R. P. W. Duin, and M. Loog, "Automated colorectal cancer diagnosis for whole-slice histopathology," *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 7512, pp. 550–557, 2012.

[73] B. Karaçali and T. Tözeren, "Automated detection of regions of interest for tissue microarray experiments: an image texture analysis," *BMC Medical Imaging*, vol. 7, 2007.

[74] A. M. Khan, H. El-Daly, and N. M. Rajpoot, "A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images," *Journal of Pathology Informatics*, vol. 4, 2013.

[75] A. M. Khan, H. El-Daly, E. Simmons, and N. M. Rajpoot, "HyMaP: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images," *Journal of Pathology Informatics*, vol. 30, 2013.

[76] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 1729–1738, 2014.

[77] E. Kim, H. Li, and X. Huang, "A hierarchical image clustering cosegmentation framework," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, 2012, pp. 686–693.

[78] J. Kononen, L. Bubendorf, A. Kallioniemi, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, and O.-P. Kallioniemi, "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nature Medicine*, vol. 4, pp. 844–847, 1998.

[79] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *IEEE Conference on Computer Vision (ICCV)*, Beijing, China, 2005, pp. 1284–1291.

[80] M. Kuse, T. Sharma, and S. Gupta, "A Classification Scheme for Lymphocyte Segmentation in H&E Stained Histology Images," in *Lecture Notes in Computer Science*, D. Ünay, Z. Çataltepe, and S. Aksoy, Eds. Springer Berlin / Heidelberg, 2010, vol. 6388, ch. 24.

[81] S. Lackhani, S. A. Dilly, and C. J. Finlayson, *Basic Pathology*. Hodder Arnold, 2009.

[82] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27, no. 8, pp. 1265–1278, 2005.

[83] ——, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, U.S., 2006, pp. 2169–2178.

[84] Leica Microsystems. (2012, November) Genie. [Online]. Available: http://www.leicabiosystems.com/pathology-imaging/aperio-epathology/analyze/details/product/genie

[85] ——. (2012, August) Leica Tissue IA. [Online]. Available: http://www.leica-microsystems.com/fileadmin/downloads/Tissue%20IA/Brochures/Tissue_IA-Brochure_en.pdf

[86] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "TurboPixels: Fast superpixels using geometric flows," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31, no. 12, 2009.

[87] Q. Li, C. Yao, L. Wang, and Z. Tu, "Randomness and sparsity induced codebook learning with application to cancer image classification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Providence, Rhode Island, 2012, pp. 16–23.

[88] W. Li, S. Liao, Q. Feng, W. Chen, and D. Shen, "Learning image context for segmentation of the prostate in CT-guided radiotherapy," *Physics in Medicine and Biology*, vol. 57, pp. 1283–1308, 2012.

[89] X. Li and H. Sahbi, "Superpixel-based object class segmentation using conditional random fields," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, 2011, pp. 1101–1104.

[90] S. Liao, G. Yaozong, J. Lian, and D. Shen, "Sparse patch-based label propagation for accurate prostate localization in CT images," *IEEE Transactions on Medical Imaging (TMI)*, vol. 32, pp. 419–434, 2013.

[91] N. Linder, J. Konsti, R. Turkki, E. Rahtu, M. Lundin, S. Nordling, C. Haglund, T. Ahonen, M. Pietikäinen, and J. Lundin, "Identification of tumor epithelium and stroma in tissue microarrays using texture analysis," *Diagnostic Pathology*, vol. 7, no. 22, 2012.

[92] A. Madabushi, "Digital pathology image analysis: opportunities and challenges," 2009.

[93] D. Magee, D. Treanor, P. Chomphuwiset, and P. Quirke, "Context aware colour classification in digital microscopy," in *Medical Image Understanding and Analysis (MIUA)*, London, U.K., 2011.

[94] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, and P. Quirke, "Colour normalisation in digital histopathology images," in *MICCAI Workshop on Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy*, London, U.K., 2009, pp. 100–111.

[95] S. Mayor, "A fifth of women with breast cancer have a recurrence, new UK figures show," *British Medical Journal*, vol. 344, 2012.

[96] E. Mercan, S. Aksoy, L. G. Shaprio, D. L. Weaver, T. Brunye, and J. G. Elmore, "Localization of diagnostically relevant regions of interest in whole slide images," in *22nd International Conference on Pattern Recognition (ICPR)*, Stockholm, Sweden, 2014.

[97] H. Moch, J. Kononen, O.-P. Kallioniemi, and G. Sauter, "Tissue microarrays: what will they bring to molecular and anatomic pathology?" *Advances in Anatomic Pathology*, vol. 8, pp. 14–20, 2001.

[98] Z. M. A. Mohammed, D. C. McMillan, B. Elsberger, J. J. Going, C. Orange, E. Mallon, J. C. Doughty, and J. Edwards, "Comparison of visual and automated assessment of ki-67 proliferative activity and their impact on outcome in

primary operable invasive ductal breast cancer," *British Journal of Cancer*, vol. 106, 2012.

[99] J. P. Monaco, J. E. Tomaszewski, M. D. Feldman, I. Hagemann, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi, "High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models," *Medical Image Analysis*, vol. 14, pp. 617–629, 2010.

[100] A. Montillo, J. Shotton, J. Winn, J. Iglesias, D. Metaxas, and A. Criminisi, "Entangled decision forests and their application for semantic segmentation of CT images," in *Information Processing in Medical Imaging*. Springer, 2011, pp. 184–196.

[101] D. Munoz, J. A. Bagnell, and M. Hebert, "Stacked hierarchical labeling," in *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, Crete, Greece, 2010, pp. 57–70.

[102] S. Naik, S. Doyle, S. Agner, A. Madabhushi, F. M., and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Paris, France, 2008, pp. 284 –287.

[103] S. K. Nath, F. Bunyak, and K. Palaniappan, "Robust tracking of migrating cells using four-color level set segmentation," *Advanced Concepts for Intelligent Vision Systems*, vol. 4179, pp. 920–932, 2006.

[104] N. Nayak, H. Chang, A. Borowsky, P. Spellman, and B. Parvin, "Classification of tumor histopathology via sparse feature learning," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, San Francisco, U.S., 2013, pp. 410–413.

[105] P. Neubert and P. Protzel, "Superpixel benchmark and comparison," in *Forum Bildverarbeitung*, Germany, 2012.

[106] K. Nguyen, B. Sabata, and A. K. Jain, "Prostate cancer grading: Gland segmentation and structural features," *Pattern Recognition Letters*, vol. 33, pp. 951–961, 2012.

[107] J. Ni, M. K. Singh, and C. Bahlmann, "Fast radial symmetry detection under affine transformations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, 2012, pp. 932–939.

[108] Oncomark. (2014, October) Automated Quantitation of IHC Scoring. [Online]. Available: http://www.oncomark.com/go/products/cell-mark

[109] G. Orchard and B. Nation, *Histopathology*, ser. Foundations of Biomedical Science. Oxford University Press, 2011.

[110] C. Panagiotakis, E. Ramasso, and G. Tziritas, "Lymphocyte segmentation using the transferable belief model," in *International Conference on Recognizing Patterns in Signals, Speech, Images, and Videos*, Istanbul, Turkey, 2010, pp. 253–262.

[111] R. L. Parker, D. G. Huntsman, D. W. Lesack, J. B. Cupples, D. R. Grant, M. Akbari, and C. B. Gilks, "Assessment of interlaboratory variation in the immunohistochemical determination of estrogen receptor status using a breast cancer tissue microarray," *Americal Journal of Clinical Pathology*, vol. 117, pp. 723–728, 2002.

[112] M. Parsons and H. Grabsch, "How to make tissue microarrays," *Diagnostic Histopathology*, vol. 15, pp. 142–150, 2009.

[113] PathXL. (2014, December) PathXL TMA. [Online]. Available: http://www.pathxl.com/pathxl-research/pathxl-tma

[114] ——. (2014, November) TissueMark. [Online]. Available: http://www.pathxl.com/files/brochures/TissueMark-PathXL.pdf

[115] Y. Peng, Y. Jiang, L. Eisengart, M. A. Healy, F. H. Straus, and X. J. Yang, "Computer-aided identification of prostatic adenocarcinoma: Segmentation of glandular structures," *Journal of Pathology Informatics*, vol. 2, 2011.

[116] S. E. Pinder, J. P. Brown, C. Gillett, C. A. Purdie, V. Speirs, A. M. Thompson, A. M. Shaaban, and on behalf of the Translational Subgroup of the NCRI Breast Clinical Studies Group, "The manufacture and assessment of tissue microarrays: suggestions and criteria for analysis, with breast cancer as an example," *Journal of Clinical Pathology*, vol. 66, pp. 169–177, 2013.

[117] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[118] I. Poole, "Optimal probabilistic relaxation labeling," in *Proceedings of the British Machine Vision Conference (BMVC)*, Oxford, U.K., 1990.

[119] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Gaussian mixture models and k-means clustering," in *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3rd ed. New York, U.S.: Cambridge University Press, 2007.

[120] C. A. Purdie, L. Baker, A. Ashfield, S. Chatterjee, L. B. Jordan, P. Quinlan, D. J. A. Adamson, J. A. Dewar, and A. M. Thompson, "Increased mortality in HER2 positive, oestrogen receptor positive invasive breast cancer: a population-based study," *British Journal of Cancer*, vol. 103, pp. 475–481, 2010.

[121] X. Qi, F. Xing, D. J. Foran, and L. Yang, "Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set," *IEEE Transaction on Biomedical Engineering*, vol. 59, pp. 754–765, 2012.

[122] J. A. Ramas-Vara and M. A. Miller, "When tissue antigens and antibodies get along: Revisiting the technical aspects of immunohistochemistry - the red, brown, and blue technique," *Veterinary Pathology*, vol. 51, pp. 42–87, 2014.

[123] A. E. Rizzardi, A. T. Johnson, R. I. Vogel, S. E. Pambuccian, J. Henriksen, A. P. N. Skubitz, G. J. Metzger, and S. C. Schmechel, "Quantitative comparison

of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring," *Diagnostic Pathology*, vol. 7, no. 42, 2012.

[124] G. K. Rohde, J. A. Ozelek, A. V. Parwani, and L. Pantanowitz, "Carnegie Mellon University bioimaging day 2014: Challenges and opportunities in digital pathology," *Journal of Pathology Informatics*, vol. 32, no. 5, 2014.

[125] A. Sapino, C. Marchió, R. Senetta, I. Castellano, L. Macri, P. Cassoni, G. Ghisolfi, M. Cerrato, E. D'Ambrosio, and G. Bussolati, "Routine assessment of prognostic factors in breast cancer using a multicore tissue microarray procedure," *Virchows Archiv*, vol. 449, pp. 288–296, 2006.

[126] C. Schmid and R. Mohr, "Matching by local invariants," INRIA, Tech. Rep. RR-2644, 1995.

[127] P. J. Schüffler, T. J. Fuchs, C. S. Ong, P. J. Wild, N. J. Rupp, and J. M. Buhmann, "TMARKER: A free software toolkit for histopathological cell counting and staining estimation," *Journal of Pathology Informatics*, vol. 4, 2013.

[128] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 22, pp. 888–905, 2000.

[129] G. Shu, A. Dehghan, and M. Shah, "Improving an object detector and extracting regions using superpixels," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, U.S., 2013, pp. 3721–3727.

[130] K. Sirinukunwattana, D. R. J. Snead, and N. M. Rajpoot, "A novel texture descriptor for detection of glandular structures in colon histology images," *Proc. SPIE*, vol. 9420, 2015.

[131] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pp. 591–605, 2009.

[132] B. W. Stewart and C. P. Wild, Eds., *World Cancer Report 2014*. IARC Press, 2014.

[133] J. Thiran and B. Macq, "Morphological feature extraction for the classification of digital images of cancerous tissues," *IEEE Transactions on Biomedical Engineering*, vol. 43, pp. 1011–1020, 1996.

[134] T. A. Thomson, C. Zhou, C. Chu, and B. Knight, "Tissue microarray for routine analysis of breast biomarkers in the clinical laboratory," *American Journal of Clinical Pathology*, vol. 132, pp. 899–905, 2009.

[135] S. Tse, L. Bradbury, J. W. L. Wan, H. Djambazian, R. Sladek, and T. Hudson, "A combined watershed and level set method for segmentation of brightfield cell images," *Medical Imaging 2009: Image Processing*, vol. 7259, 2009.

[136] Z. Tu, "Auto-context and its application to high-level vision tasks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[137] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, pp. 1744–1757, 2010.

[138] D. Turbin, S. Leung, M. Cheang, H. Kennecke, K. Montgomery, S. McKinney, D. Treaba, N. Boyd, L. Goldstein, S. Badve, A. M. Gown, M. van de Rijn, T. O. Nielson, C. B. Gilks, and D. G. Huntsman, "Automated quantitative analysis of estrogen receptor expression in breast carcinoma does not differ from expert pathologist scoring: a tissue microarray study of 3,484 cases," *Breast Cancer Research and Treatment*, vol. 110, no. 3, pp. 417–426, 2008.

[139] M. van den Bergh, X. Boix, G. Roig, B. de Capitani, and L. van Gool, "SEEDS: Superpixels extracted via energy-driven sampling," in *Proceedings of the 12th European Conference on Computer Vision (ECCV)*, Florence, Italy, 2012.

[140] V. J. van Diest, P. van Dam, S. C. Henzen-Logmans, E. Berns, M. E. van der Burg, and I. Vergote, "A scoring system for immunohistochemical staining: consensus report of the task force for basic research of the EORTC-GCCG," *Journal of Clinical Pathology*, vol. 50, pp. 801–804, 1997.

[141] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," *10th European Conference on Computer Vision (ECCV)*, pp. 705–718, 2008.

[142] M. Veta, P. J. van Diest, S. M. Willems, H. Wang, A. Madabhushi, A. Cruz-Roa, F. Gonzalez, A. B. L. Larsen, J. S. Vestergaard, A. B. Dahl, D. C. Cireşan, J. Schmidhuber, A. Giusti, L. M. Gambardella, F. Boray Tek, T. Walter, C.-W. Wang, S. Kondo, B. J. Matuszewski, F. Precioso, V. Snell, J. Kittler, T. E. de Campos, A. M. Khan, N. M. Rajpoot, E. Arkoumani, M. M. Lacle, M. A. Viergever, and J. P. W. Pluim, "Assessment of algorithms for mitosis detection in breast cancer histopathology images," *Medical Image Analysis*, vol. 20, pp. 237–248, 2015.

[143] G. Viale, "The current state of breast cancer classification," *Annals of Oncology*, vol. 23, pp. x207–x210, 2012.

[144] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 13, pp. 582–598, 1991.

[145] D. Voduc, C. Kenney, and T. O. Nielsen, "Tissue microarrays in clinical oncology," *Seminars in Radiation Oncology*, vol. 18, pp. 89–97, 2008.

[146] W. H. Wan, M. B. Fortuna, and P. Furmanski, "A rapid and efficient method for testing immunohistochemical reactivity of monoclonal antibodies against multiple tissue samples simultaneously," *Journal of Immunological Methods*, vol. 103, pp. 121–129, 1987.

[147] C.-W. Wang, D. Fennell, I. Paul, K. Savage, and P. Hamilton, "Robust automated tumour segmentation on histological and immunohistochemical tissue images," *PLoS ONE*, vol. 6, 2011.

[148] X. Wang, X. Bai, W. Liu, and L. J. Latecki, "Feature context for image classification and object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado, U.S., 2011, pp. 961–968.

[149] B. Weyn, G. van de Wouwer, A. van Daele, P. Scheunders, D. van Dyck, E. van Marck, and W. Jacob, "Automated breast tumor diagnosis and grading based on wavelet chromatin texture description," *Cytometry*, vol. 33, pp. 32–40, 1998.

[150] A. Wiliem, C. Sanderson, Y. Wong, P. Hobson, R. F. Minchin, and B. C. Lovell, "Automatic classification of human epithelial type 2 cell indirect immunofluorescence images using cell pyramid matching," *Pattern Recognition*, 2013.

[151] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.

[152] X. Wu and S. K. Shah, "Cell segmentation in multispectral images using level sets with priors for accurate shape recovery," in *International Symposium on Biomedical Imaging: From Nano to Macro*, Chicago, U.S., 2011, pp. 2117–2120.

[153] J. Xu, A. Janowczyk, S. Chandran, and A. Madabhushi, "A high-throughput active contour scheme for segmentation of histopathological imagery," *Medical Image Analysis*, vol. 15, pp. 851–862, 2011.

[154] Y. Xu, J. Zhang, E. I. Chang, M. Lai, and Z. Tu, "Context-constrained multiple instance learning for histopathology image segmentation," *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 7512, pp. 623–630, 2012.

[155] W. Yu, H. Lee, S. Hariharan, W. Bu, and S. Ahmed, "Level set segmentation of cellular images based on topological dependence," *Advances in Visual Computing*, vol. 5358, pp. 540–551, 2008.

[156] B. Zhang, C. Zimmer, and J.-C. Olivo-Marin, "Tracking fluorescent cells with coupled geometric active contours," in *IEEE International Symposium on Biomedical Imaging: Nano to Macro*, Virginia, U.S., 2004, pp. 476–479.

[157] Y. Zhang, B. Zhang, F. Coenen, and W. Lu, "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier emsembles," *Machine Vision and Applications*, vol. 24, pp. 1405–1420, 2013.

[158] Y. Zheng, H. Lu, C. Jin, and X. Xue, “Incorporating spatial correlogram into bag-of-features model for scene categorization,” *Asian Conference on Computer Vision (ACCV)*, vol. 5994, pp. 333–342, 2010.